Seeing things as people: anthropomorphism and common-sense psychology

Thesis

Seeing things as people

Anthropomorphism and common-sense psychology

Stuart Neil Kennaway Watt

Thesis submitted for the degree of PhD in Cognitive Science (Psychology Discipline)

23rd January 1997

RESEARCH DEGREES CENTRE

LIBRARY AUTHORISATION FORM

Please return this form to the The Research Degrees Centre with the two bound copies of your thesis to be deposited with the University Library.
All students should complete Part 1. Part 2 only applies to PhD students.

Student: STUART WATT PI: M7121466

Degree: PhD

Thesis title: SEEING THINGS AS PEOPLE: ANTHROPOMORPHISM AND COMMON-SENSE PSYCHOLOGY

Part 1 Open University Library Authorisation [to be completed by all students]

I confirm that I am willing for my thesis to be made available to readers by the Open University Library, and that it may be photocopied, subject to the discretion of the Librarian.

Signed: Stuart N.L. Watt Date: 23rd April 1998

Part 2 British Library Authorisation [to be completed by PhD students only]

If you want a copy of your PhD thesis to be available on loan to the British Library Thesis Service as and when it is requested, you must sign a British Library Doctoral Thesis Agreement Form. Please return it to the Research Degrees Centre with this form. The British Library will publicise the details of your thesis and may request a copy on loan from the University Library. Information on the presentation of the thesis is given in the Agreement Form.

Please note the British Library have requested that theses should be printed on one side only to enable them to produce a clear microfilm. The Open University Library sends the fully bound copy of theses to the British Library.

The University has agreed that your participation in the British Library Thesis Service should be voluntary. Please tick either (a) or (b) to indicate your intentions.

[a] ☑ I am willing for the Open University to loan the British Library a copy of my thesis. A signed Agreement Form is attached.

[b] ☐ I do not wish the Open University to loan the British Library a copy of my thesis.

Signed: Stuart N.L. Watt Date: 23rd April 1998.

Abstract

This thesis is about common-sense psychology and its role in cognitive science. Put simply, the argument is that common-sense psychology is important because it offers clues to some complex problems in cognitive science, and because common-sense psychology has significant effects on our intuitions, both in science and on an everyday level.

The thesis develops a theory of anthropomorphism in common-sense psychology. Anthropomorphism, the natural human tendency to ascribe human characteristics (and especially human mental characteristics) to things that aren't human, is an important theme in the thesis. Anthropomorphism reveals an endemic anthropocentricity that deeply influences our thinking about other minds. The thesis then constructs a descriptive model of anthropomorphism in common-sense psychology, and uses it to analyse two studies of the ascription of mental states. The first, Baron-Cohen *et al.*'s (1985) false belief test, shows how cognitive modelling can be used to compare different theories of common-sense psychology. The second study, Searle's (1980) 'Chinese Room', shows that this same model can reproduce the patterns of scientific intuitions taken to systems which pass the Turing test (Turing, 1950), suggesting that it is best seen as a common-sense test for a mind, not a scientific one. Finally, the thesis argues that scientific theories involving the ascription of mentality through a model or a metaphor are partly dependent on each individual scientist's common-sense psychology.

To conclude, this thesis develops an interdisciplinary study of common-sense psychology and shows that its effects are more wide ranging than is commonly thought. This means that it affects science more than might be expected, but that careful study can help us to become mindful of these effects. Within this new framework, a proper understanding of common-sense psychology could lay important new foundations for the future of cognitive science.

Contents

Acknowledgements

To my parents and family, with all my love.

To my friends and colleagues for their help and support—including John Domingue, Jenny and Darryl Gove, Simon Masterton, Enrico Motta, Paul Mulholland, Susan Stuart and Norman Gray, Arthur Stutt, and Zdenek Zdrahal—you are all wonderful people. Special thanks to Susan and Arthur for theoretical and literary help beyond the call of duty.

To George Kiss, who, as supervisor, helped me disentangle the threads in this thesis. and to Jill Cohen, also as supervisor, who helped me tie them back together again.

To Marc Eisenstadt, for being generally wonderful, and for making HCRL and KMI fun places to hang out.

To Martin Stephenson, for writing 'There Comes A Time'.

To various cats—to Morag and Mungo for being both beautiful and fun; and to Macgregor for some hints on chapter 7.

To Chris McKillop, for being there for me, and for her love. And for not putting 'kittens', or 'dinosaurs' anywhere in this thesis.

Finally, to three people who changed my life in different ways, but who never lived to see just how much. I owe them a dept that I could never have repaid. This thesis is for them—Anne Drummond, Hank Kahney, and Al Roth.

A road map

Because this thesis is interdisciplinary, it may be helpful to think about its organisation with this diagram. Some of the contributing disciplines are shown vertically, with a typical cognitive science methodological structure shown horizontally. When a thesis chapter adds to a discipline's perspective on an element of this methodology, it is shown in the box at the intersection between the two.

As this road map shows, the thesis generally follows a kind of 'spiral' pattern through the disciplines, as it moves through the background, the models, and a discussion of the issues, to unfold its argument.

		Background	Method	Results	Discussion
Discipline	Philosophy	Chapters 2 and 5	Chapters 6, 7, and 11	Chapter 12	Chapters 13, 14, and 15
	Psychology	Chapter 3	Chapters 6, 7, 8, and 9	Chapter 10	Chapters 14 and 15
	Artificial intelligence	Chapters 4 and 5	Chapters 6 and 11	Chapter 12	Chapters 13, 14, and 15

Publication details

Material in the following chapters of this thesis has been published:

A revised version of chapter 6, 'A common-sense psychology manifesto', has been published as "A Brief Naive Psychology Manifesto" in *Informatica, 19*(4), 1995, pp. 495-500, a special issue on the differences between minds and computers.

A revised version of chapter 13, 'Common-sense psychology and the inverted Turing test', has been published as "Naive Psychology and the Inverse Turing Test", *PSYCOLOQUY, 7*(14), 1997.

A revised version of the first part of chapter 14, 'Methodological implications', has been published as "The Lion, the Bat, and the Wardrobe: Myths and Metaphors in Cognitive Science", S. O Nuálláin, P. McKevitt, and E. Mac Aogáin (eds.), *Two Sciences of Mind: Readings in Cognitive Science and Consciousness* (pp. 51-61), Amsterdam: John Benjamins, 1997.

It is here that common-sense psychology begins to play an important role. Briefly, common-sense psychology is the natural human faculty that enables us, from childhood, to think about our own and other people's minds. The philosophical problem of whether we can know about other minds in principle is replaced by the psychological problem of how we know about other minds in practice. The price that we pay is a high one; we lose sight of the philosophical problem of whether or not something actually has a mind. But in return we can look at the intuitions which, from the evidence available, help us decide whether or not we think that something has a mind.

In this thesis I will carry out a far more detailed analysis of that part of common-sense psychology which is concerned with how people decide whether or not something has a mind. This is an important issue for cognitive science, because it shows that many of the apparently philosophical criteria of mindedness are actually rooted in ordinary human common-sense psychological intuitions. There is a two way connection between science and common-sense psychology; first, science can be used to study common-sense psychology, as I will do in this thesis, but second, common-sense psychology underpins certain parts of science, as I will show in this thesis. We can use science to study the effects of common-sense psychology on science, and if we don't, eventually we may find that some of our scientific assumptions depend on unfounded intuitions.

This thesis will argue that one aspect of common-sense psychology (which, for want of a better word, I will call 'anthropomorphism') plays a fundamental role in our discrimination between things with and without minds. Anthropomorphism as a psychological phenomenon has never really been deeply studied, but the studies which have been carried out all show that there are some extremely strong effects involved. The correlations between people's ascribed similarity, and their ascriptions of cognitive competence and similar experience, are all very close. Among other things, this phenomenon is anthropocentric, intuitive, sensitive to individual differences, and plays an important role both in everyday and scientific judgements about minds. The implications of this run deep in philosophy, psychology, and artificial intelligence. Particularly and especially, it suggests that significant parts of cognitive science theory are dependent, in practice, on these intuitions, and while this does not necessarily weaken them, it does change their nature in ways that might have significant implications for the scientific method in this discipline.

So why is a study of common-sense psychology important? The answer I would like to give is that a proper cognitive science is impossible without it, both methodologically and theoretically. This is because people only see minds from the point of view of their natural human common-sense psychology, and this is reflected both in our scientific and our everyday judgements of each other, of theoretical models in cognitive science, and of animal and machine intelligence. Without a proper understanding of the effects of this common-sense psychology, it can become impossible to separate the psychology of the observer from the psychology of the observed. A study of common-sense psychology is important because it gives us a window onto the uniqueness of our human point of view, and can help us become mindful of it.

Different kinds of common-sense psychology

One of the problems with studying common-sense psychology is that it has a number of different names, depending on which discipline happens to be interested in it. Unfortunately, because of the differing discipline perspectives, these terms are not completely synonymous. It is worth, therefore, briefly reviewing these terms and their different perspectives.

Folk psychology; also known as belief-desire psychology. Originally, the term 'folk psychology' was introduced by Wundt (1900-1920), who was, perhaps, the first to study the psychology of common-sense psychology in science. More recently, the term has become the favoured term for common-sense psychology in philosophy. I have not used the term 'folk psychology' for two reasons. Like Wilkes (1991) I think it sounds too "folksy" and therefore perhaps obscures the sheer magnitude of common-sense psychology. Secondly, though, and perhaps more importantly, there is a wide range of conceptually different interpretations of the term. Some, for example, interpret it as being rigidly constructed of beliefs and desires as sentences in the head with genuine causal powers, where others take it far more loosely, as a mental faculty for working with other people's mental states. To avoid these additional connotations, I will not generally use this term.

Naive psychology. Again, the term has an origin different from its current use. It was originally coined by Heider (1958), and referred to common-sense psychology in a loose sense, but directly contrasted with scientific psychology. Today, the term 'naive psychology' is perhaps already obso-

lete. Its heyday was in the field of artificial intelligence in the mid 1980s, when common-sense reasoning was beginning to address psychology as well as physics; that is, the term was intended to draw comparisons with common-sense—or naive—physics.

There are also good reasons for not using the term 'naive psychology'. When naive physics became the vogue in artificial intelligence research, it was generally coupled with a particular methodology—the methodology of using sentences in logic to describe common-sense reasoning, but without any description of how the use of these sentences was controlled. Both the presence of logic (as a causal description of common sense) and the absence of control have been at the heart of the criticisms of common-sense reasoning in artificial intelligence—and therefore of naive psychology. I am avoiding the term 'naive psychology', then, to distance myself from these methodological assumptions.

Theory of mind; also known as mindreading (Whiten, 1991), natural psychology (Humphrey, 1976). The term 'theory of mind' was first used by Premack and Woodruff (1978), but since this work in comparative psychology, it has become one of the favoured terms for common-sense psychology in psychology as a whole. 'Theory of mind' is perhaps the most problematic candidate term; here the problems lie in the word 'theory'. For some, the word 'theory' means a *scientific* theory like Newton's physics, made up of laws and rules of behaviour, and rather similar in structure to the logical models of common-sense psychology developed and explored in artificial intelligence. For others, 'theory' just means a body of tricks that can be used to guess at and predict another person's behaviour. For the most part, then, I will steer well clear of the word 'theory' in this thesis.

For these reasons, and to avoid getting bogged down in a terminological swamp, I will use the term 'common-sense psychology' throughout most of this thesis, except where I want to draw attention to one particular interpretation. This term has the advantage of avoiding the unintended interpretations I have discussed, while remaining acceptable in all the disciplines of philosophy, psychology, and artificial intelligence.

Common-sense psychology, then, has different facets in several different disciplines. Because of this, the research presented in this thesis is interdisciplinary, and combines different aspects of the three principal areas involved in this part of cognitive science; philosophy, psychology, and artificial intelligence. I will briefly present the thesis' main contributions to these three areas in the next three sections, and I will come back to discuss these contributions more substantially in chapter 15.

Contributions: common-sense psychology in philosophy

To begin with, in philosophy, there are intimate connections between several long running philosophical problems and common-sense psychology. The 'mind body' problem and the 'other minds' problem are perhaps the best known of these, but such is the interconnectivity of philosophy that almost every aspect of philosophy is needed to deal with human common sense in one way or another. McCarthy even suggested that "philosophy could be defined as an attempted *science of common sense*" (McCarthy, 1979, original emphasis). Perhaps this thesis' most important contribution to philosophy is in the other minds problem. In particular, I will suggest that this is not only, and perhaps even not principally, a philosophical problem; the theories in this thesis show that the problem becomes a psychological one, one which people solve through common sense psychology. As Searle says "except when doing philosophy, there really is no 'problem' about other minds" (Searle, 1992). The question becomes one of epistemology rather than ontology: we can study how people know about other minds rather than worrying about whether it is possible to know about other minds in principle.

The 'other minds' problem, then, is simply this: how do you know whether somebody else has a mind? Well, in practice, depending on the circumstances, you'll see them either as psychological objects or as physical objects. If you see them as psychological to the point where you ascribe them subjectivity, their own experience of the world, you will, pending contrary evidence, say that they really do have a mind. And this is where anthropomorphism begins to play an important role. As I'll show in chapters 7 and 8, the factors which influence this at a psychological level are the factors of anthropomorphism. The 'mind-body' problem is perhaps more opaque than the

'other minds' problem; it is complicated by functionalist and other kinds of dressing, which don't really help us to understand the problem. Again, though, in practice, people look at the form and character of the body, and are influenced fundamentally in their ascription of mentality by these, through the same anthropomorphic effects.

The stance I will take here is that common-sense psychology is endemic; that is, that there is no way that it can ever be eliminated from cognitive science in practice. That is, the theory of 'eliminative materialism' (Churchland, 1981; Stich, 1983)—or simply 'eliminativism'—is simply false. Eliminative materialism argues that the elements which make up common-sense psychology, such as beliefs and desires: "are like phlogiston, caloric, and witches; they are the mistaken posits of a radically false theory" (Stich & Nichols, 1992), and for this reason, common-sense psychology can have no possible role in science.

Although I'd agree with Clark's (1989) criticism of eliminative materialism—that theories of common-sense psychology are not the same kind of thing as theories of science, so they cannot be false in the same way—this isn't the main issue here. In this thesis, I take the complementary position that an understanding of common-sense psychology is necessary for us as scientists to be aware of the real meanings of our models, metaphors, and architectures. There is, in effect, no way to create the perfect neuroscience desired by eliminative materialists until we can distinguish between a real neuroscience and the ghosts in the network that are artifacts of our common-sense psychology. To do this, we need to study common-sense psychology. I will return to these methodological issues in chapter 14.

Finally, another philosophical contribution of this thesis is to show that there is a close affinity between Dennett's (1971) model of intentional systems, and research in psychology (e.g. that of Carey, 1985) which breaks down people's common-sense reasoning in a very similar manner. There is a close correspondence between the philosophical and psychological aspects of the theories, even though Dennett's theory was intended principally as a philosophical solution to issues such as the mind-body problem. The gap between the different theories is, in fact, in Dennett's (1971) "rationality assumption". In this thesis, I'll show that the part of the rationality assumption can be played by the rather under-investigated phenomenon of anthropomorphism, which does indeed

seem to be a disposition to take the intentional stance. This influence on when people take the intentional stance, as opposed, say, to the physical stance, seems to be a missing factor in many of the different theories of common-sense psychology. In the model I'll develop in this thesis, therefore, I construct a far more complete model of anthropomorphism and connect this into faculties of common-sense physics and common-sense psychology. It is this model which is at the core of this thesis, and which is used to investigate the effects of common-sense psychology in more detail.

But the effects of common-sense psychology go beyond this. Another significant theme of this thesis is the point that people take the intentional stance not only to physical objects and psychological agents, but to abstract theoretical objects (such as psychological models). I will discuss the case for common-sense psychology playing a crucial role in people's intuitions about models later in the thesis. Briefly, I'll show that different metaphors and models make a difference, over and above their explanatory power, in the reactions they invoke from our own common-sense psychology. As the model in chapters 11 and 12 will show, the patterns of these intuitions can be predictable, and while individual differences also play a substantial role, within limits the range of these differences can also be predicted. The model of common-sense psychology in this thesis does seem able to predict, at least partly, people's intuitive reactions to different psychological models.

Contributions: common-sense psychology in psychology

In psychology very different problems arise. Common-sense psychology is already an area of active research in developmental psychology (e.g. Baron-Cohen *et al.*, 1985) and comparative psychology (e.g. Gómez *et al.*, 1993) in particular, but methodologically it is working with one hand tied behind its back. It badly needs models which are better specified, so that the different theories can be compared with each other and to human subjects. This means borrowing the approach of computational modelling for common-sense reasoning. This methodology has already offered substantial benefit in developing models for other aspects of psychology, but so far com-

mon-sense psychology has escaped almost completely from this methodological approach. An important goal of this thesis is to increase the scope of computational modelling in common-sense psychology.

The stance I will take here is that common-sense psychology is natural; that is, that it has evolved through natural selection and therefore is related to, although clearly not the same as, common-sense psychology in other animals. I have not tried to develop a complete new theory for common-sense psychology within this thesis; there are enough of them already in the existing psychological and philosophical literatures (e.g. Gordon, 1986; Perner, 1991; Wellman, 1990). Instead, I will develop a modelling framework which allows the different theories to be compared. This is traditional computational modelling in the manner of Newell (1992). The contribution here, then, is that the thesis offers a new tool to help analyse and compare existing theories, rather than offering a new theory or synthesis.

Perhaps the biggest single contribution of the thesis in this area, though, is the model of anthropomorphism described in chapters 6, 7, and 8. While still fairly immature, this is an area which has largely been ignored by psychologists, yet which is potentially immensely important for a proper understanding and approach to science in other fields, ethology for example (Kennedy, 1992). Here, although there is a fairly good evidential basis for some of the factors involved, others are still only known from rather patchy evidence. These factors need fuller psychological study to look at the nature of anthropomorphism in more detail. But, here again, the modelling technique has proved useful, in that it has provided a clear hypothetical structure for the phenomenon of anthropomorphism, one which can now be taken as a strong hypothesis to be investigated experimentally.

Finally, another significant contribution in this field is to throw new light onto the distinction between the two principal competing theories of common-sense psychology, the 'theory theory' and the 'simulation theory'. The 'theory theory' proposes that common-sense psychology is organised as a body of knowledge, structured in law-like ways which have some affinity to a scientific theory, while the 'simulation theory' argues that people's common-sense interpretation of other people's behaviour is made through role taking and introspection, by putting oneself 'in the

other person's shoes'. I will review these theories in more detail in chapter 3, and show through the model in chapter 10 that, as argued by Davies (1994), the two can behave identically; that is, that the apparent egocentricity of the simulation theory may indeed be only apparent, and that the functional effect of a simulation may be identical to that of a theory.

Contributions: common-sense psychology in artificial intelligence

In artificial intelligence, common-sense reasoning has been stuck into deductive modes of reasoning for so long that more recent reactions against the use of logic have perhaps gone too far. In this discipline, the thesis will bring a new element of psychological plausibility to common-sense psychology, and extend it to deal with a number of psychological phenomena not normally addressed in common-sense reasoning in artificial intelligence.

An important contribution here is a re-evaluation of the Turing test (Turing, 1950). I'll introduce and describe a modification of the Turing test, interpreting it as a framework for assessing social interaction rather than intelligence. This interpretation of the Turing test offers a number of insights: first, it clarifies some of the standard misinterpretations of and biases in the test; second, some of these misinterpretations and biases offer important clues to the nature of common-sense psychology; and third, the inverted Turing test I propose in chapter 13 is useful in analysing an agent's competence at the ascription of mental states to others. It is Searle's (1980) "Chinese Room" argument that makes the role of intuition in the Turing test especially clear, because Searle's version says more about the observer's knowledge than Turing's original, and shows how completely the ascription of intelligence is bound up with this knowledge. This needn't mean that the Turing test should be abandoned as a tool for research in artificial intelligence—indeed, I suggest that it can be particularly helpful as a tool for studying the common-sense intuitions that shape the ascription of intelligence. Accordingly, passing or failing the Turing test, even in a simple version such as Loebner's competition, may not be able to tell us about the nature of intelligence, but might tell us what it takes to make people ascribe intelligence, even to something which isn't a human.

The stance I will take here is simply that common-sense psychology is important. This is all too easily forgotten. I present a kind of 'manifesto' for this in chapter 6. The rather rigid and formal models of common-sense psychology which are common in artificial intelligence (e.g. Cohen & Levesque, 1990; Moore, 1985; Shoham, 1992) recognise the usefulness of it, but not really its importance. In this light, I argue that artificial intelligence has been guilty of a worse version of anthropocentricity than even philosophy or psychology. This is ironic, because of all these disciplines, it is perhaps the one that has adopted most animal metaphors and principles (e.g. Brooks, 1991a). Perhaps this was unwise; in artificial intelligence, there has been a persistent view that there might be such a thing as intelligence which could live outside human psychology, a view which I call the 'alien intelligence hypothesis'. This has been adopted by, among others, McCarthy (1979), French (1990), Hayes (1985b), and Turing (1950). The problem with this hypothetical alien intelligence is that if there is a kind of intelligence which is outside human psychology, how do we recognise it as intelligence, when it seems that human (common-sense) psychology is, effectively, what we use to recognise intelligence.

This implies that the alien intelligence hypothesis is simply false, and this is a significant point in this thesis. I don't want to get into big terminological pot-holes here; the intention behind this argument is merely to suggest that the only behaviours that can be called intelligent, are those which we humans recognise as intelligent. There can never really be, therefore, any kind of intelligence that exists outside our humanity. I will discuss this argument a lot more fully in chapters 6 and 13, and I don't want to add much to those points here. There are several arguments which point to this conclusion, including French's (1990) argument about the inseparability of the cognitive and subcognitive levels, as well as this thesis' own argument from an inverted version of the Turing test in chapter 13. However, where French argues that the Turing test can't identify intelligence in general, and is therefore flawed, I will argue that no other kind of intelligence is possible and since the Turing test can only identify human intelligence, it is mostly valid. But either way, should the alien intelligence hypothesis be false, important research themes in artificial intelligence are significantly undermined.

On the importance of anthropomorphism

Much of this thesis centres around the concept of anthropomorphism and its importance as a central aspect of common-sense psychology. This can be a confusing term, so it is especially important to be clear about its use. I do not want to be read as suggesting that common-sense psychology is in any way structured by the 'jokey' kind of anthropomorphism that appears in, for example, the works of Aesop, Lewis Carroll, or Rev. W. D. Awdry. It is not 'Thomas the Tank Engine' psychology, although it is quite possible that the anthropomorphism in *Thomas the Tank Engine* is dependent on common-sense psychology. This research addresses serious and important issues for a science of the mind. By 'anthropomorphism' I mean the family of psychological phenomena involved in our natural, human tendency to ascribe human—and especially our own—characteristics to things which aren't specifically human. For the purposes of this thesis, the main emphasis is on people's ascription of human mental characteristics.

'Anthropomorphism' is a less than perfect term for this phenomenon, partly because of the jokiness it is often associated with. The phenomenon also includes aspects of animacy and animalness (Carey, 1985; Shultz, 1991), and after all, the Latin word *anima* means both life and soul; the concepts of having a soul, having life, and having a mind are all linguistically intertwined. In practice, animacy, in the sense of moving with intention, is also strongly implicated in the ascription of mentality to other people and things, given its correlation with anthropomorphism, especially in the natural world (Eddy, Gallup, & Povinelli, 1993). Other related phenomena include empathy and identification—these represent a kind of 'first person' anthropomorphism, where instead of ascribing human characteristics to an external animal or object, one imagines what it would be like to be that object. This is the other side of the anthropomorphic coin; instead of projecting characteristics onto an external object, one temporarily 'introjects' its characteristics into oneself. These twin aspects of anthropomorphism will be discussed in more detail in chapter 7.

So anthropomorphism is a subtle and complex mixture of phenomena; one which it will be beyond the scope of this thesis to disentangle completely. For the sake of this thesis, I am simply using the term to cover all the various psychological phenomena involved in ascribing human mental characteristics to something—whether human, animal, object, or artifact—without scien-

tific evidential warrant. This interpretation leaves the term firmly within the naturalistic and psychological domain and matches the experimental analyses of anthropomorphism reviewed in chapter 7, but, importantly, it allows us to extend it into the realms of artificial intelligence and cognitive science. This extension of anthropomorphism is at the heart of this thesis, because the implications of an apparently endemic psychological phenomenon of anthropomorphism have far reaching consequences for cognitive science and artificial intelligence.

Methodological issues

Any interdisciplinary project such as this will inevitably open a can of methodological worms. Each discipline has its own preferred set of methods, and sometimes these methods seem to be incompatible with those of other disciplines. In this project artificial intelligence techniques are used because they allow a relatively unambiguous computational model to be created—a model which can be tested and compared to human subjects. In this sense, the thesis is methodologically what Searle (1992) has called "weak psychological artificial intelligence". More of the methodological issues associated with artificial intelligence will be addressed in chapter 4 when I review common-sense reasoning in artificial intelligence, because in this discipline there are methodological assumptions with technical—and rather dubious—implications. However, following the majority of cognitive scientists, I will simply claim that artificial intelligence techniques form a useful complement—although not a replacement—to the more established approaches from philosophy and psychology (Miller, 1981; Boden, 1988).

This project will, though, stop short of Searle's "strong artificial intelligence", in that I do not want to claim that the model I will build in any sense *is* a common-sense psychology in its own right. That is, I do not pretend that this model will offer a causal account; instead, I only want to suggest that it offers a descriptive one (Clark, 1988). This distinction is an important one. However, even a descriptive model can be useful in a research programme, as it can be used to explore which hypotheses about human behaviour are more or less probable, and therefore it can contribute significantly to the evidence which invalidates a given theory.

What this thesis isn't about

Before proceeding, there are some other aspects of cognitive science that should be explicitly ruled out of the main topics of this thesis, simply to avoid confusion. This is not intended to mean that they aren't relevant or related to this thesis, or that they shouldn't form aspects of future work. I am drawing a frame around them at this point because they are more likely to add confusion than they are to assist understanding.

Consciousness. Consciousness is becoming an important part of cognitive science, and some have drawn a strong connection between it and common-sense psychology in one form or another (e.g. Humphrey, 1984; Johnson, 1988; Rorty, 1993). Consciousness is a composite of many concepts, some of which are closely related to the topics of this thesis (reflexive thought, for instance), while others are only distantly related (such as subjectivity).

On the other hand, *ascribed* consciousness is an important theme in this thesis. The whole project began as an attempt to study some of the apparently confusing regularities which show up in people's intuitions about who or what has or hasn't got consciousness. As it turns out, many aspects of consciousness are prone to misreading through common-sense psychology, and I found it necessary to work more on this misreading than on consciousness itself. There is evidence for this in Eddy *et al.*'s (1993) results, which show a close correlation between ascribed similarity and ascribed consciousness. The two distinct interpretations of Searle's (1980) 'Chinese Room' thought experiment (as about intentionality, or about consciousness) also show a distinct affinity between these two very different kinds of mental ascription; I'll discuss this later in chapter 11. So while it may not be that common-sense psychology is necessary for understanding consciousness, it is that a good theory of common-sense psychology is necessary for not *mis*understanding consciousness. If this theory of ascribed consciousness is found to be true, it is of fundamental importance to future research in cognitive science. In particular, it suggests that any attempt to study 'pure' consciousness, that is, consciousness of an objective quality, separate from an ascriber, may be fundamentally flawed by the anthropocentricity that has dogged the rest of cognitive science.

Intentionality. Intentionality is a big philosophical problem. I will raise some of its issues briefly in chapter 2, but by no means fully. It is not the goal of this thesis to provide an account of intentionality—this is a deep and fundamental problem which extends far beyond the bounds of common-sense psychology, for instance into language (Searle, 1983) and even consciousness (Searle, 1990). Although the perspectives from common-sense psychology presented in this thesis do have significant implications for intentionality, discussed in chapters 2 and 15, I would not claim that it constitutes a thorough or complete examination of the topic.

Building true artificial intelligences. I have already mentioned that this project doesn't fall into the category of strong artificial intelligence. Although many of the results of this project may be useful in carrying out stronger versions of artificial intelligence than described here—building systems which pass the modified Turing test presented in chapter 13, for example—this is, again, beyond the scope of this thesis. Instead, this thesis is narrowly focused on *human* common-sense psychology, and is perhaps better considered as being in the discipline of cognitive science rather than artificial intelligence.

These exclusions may seem rather arbitrary, but they shouldn't be considered as necessarily outside the scope of the project completely, merely that these are deep issues which cannot be fully addressed in a thesis of this scale. In all of these areas there are possibilities for future work which I will recap in the last part of the thesis; for now all I ask is a little tolerance if I fail to give a complete account of them. It may be possible to provide these accounts eventually, but in all these areas, a good account of human common-sense psychology seems to be a precondition, for one reason or another.

Overview of the thesis

This thesis is organised in four parts. This first part sets the scene for the rest of the thesis and reviews the related background material in the many disciplines which contribute to the study of common-sense psychology. The thesis is necessarily interdisciplinary, so I'll present this material as seen from the points of view of these different disciplines, all the time drawing the common threads together. This part also introduces the methodological stance for the thesis. Finally, this

part includes a review of the Turing test, showing that it is extremely sensitive to common-sense psychology, and that, by evaluating the ascription of mentality to things which are very different from people, it can reveal some important clues about what really affects this ascription.

The second part of the thesis introduces and builds a psychological theory of anthropomorphism. This is one of the thesis' main original contributions. I'll show that anthropomorphism is an important aspect of common-sense psychology, acting as a disposition to see some things as agents rather than just as objects. Here I bring together elements of the psychology of anthropomorphism with elements of Dennett's notion of different 'stances' to build a theory of how an observer can come to construe a system either as an agent or as an object, depending on its knowledge and on the context.

The third part adopts a computational modelling methodology, and transforms the theory developed in the second part into a working, although fairly shallow, cognitive model. This model is then evaluated in two reference domains; first, Baron-Cohen *et al.*'s (1985) 'false belief test', a paradigm tool in the psychology of common-sense psychology, and second, Searle's (1980) 'Chinese Room' thought experiment, which, as I'll explain, beautifully illustrates many of the effects of common-sense psychology on people's understanding of scientific models and metaphors.

The fourth and last part of the thesis returns to the theory and discusses the implications of common-sense psychology. It begins with a redescription of the theory as an 'inverted Turing test', namely, as a tool which looks at the observer's ascription of intelligence rather than the system's simulation of intelligence, as in the standard Turing test (Turing, 1950). This part of the thesis argues that, above all, common-sense psychology is pervasive to science as well as to human perceptions, and discusses the methodological implications of this. The thesis then concludes by reviewing the main theoretical contributions and implications of this research.

Summary

So this is an interdisciplinary project, and this is reflected in the thesis as a whole. There are methodological, technical, and terminological tangles which I will try to avoid, and there are related issues which I believe cannot be addressed until a more complete understanding of common-sense psychology is available. My hope is that this project makes enough contribution for a little progress in this important area. And so, without more delay, I will pass on to the next part of the thesis, which reviews the background work in these different disciplines against which these contributions will be made.

Part One

Introducing common-sense psychology

BLANK IN ORIGINAL

Chapter 2

Common-sense psychology in philosophy

Introduction to the philosophy of common-sense psychology

In the last few decades the philosophy of mind has been the target of challenges from two sources, evolutionary theory and artificial intelligence. Each has attempted to broaden the class of systems with minds, in the first case to some animals and in the second to some machines. The discussions and revisions that have happened as a result have played an important part in setting a new foundation for a science of the mind.

This has forced science to address many of the apparently simple things that humans do so naturally. One of the most important of these is common sense; people have an intuitive grasp of the behaviour of physical objects and of one another's mental states. These competences of common-sense (or 'folk') physics and common-sense (or 'folk') psychology respectively are at the heart of this new foundation for the philosophy of mind, because they are essential elements of the distinction between the physical and the mental.

But before we can start to study common-sense psychology, we need to answer a number of questions about it. We need to decide what it actually is, ontologically speaking, and this is still a very open problem. For example, Astington and Gopnik (1991) distinguish six possible alternative structures for common-sense psychology; as a theory, as a "form of life" (Wittgenstein, 1953), as an innate module, as procedural knowledge, as experience, and as a story—although these are not necessarily exclusive, and "at the risk of sounding like wishy-washy liberals, there is some level at which they must all be true" (Astington & Gopnik, 1991). I will return to a more complete study of these alternative descriptions in the next chapter, when I begin to discuss the psychology

of common-sense psychology. For now, I will turn to some of the philosophical preoccupations, namely, what constitutes common-sense psychology; that is, what structural and functional elements go to make up human common-sense psychology.

The most important constitutive elements of common-sense psychology in philosophy are a person's 'attitudes'. Different interpretations of these attitudes abound, but perhaps the most common is that a person's mental states can be represented in attitudes to a proposition about the world or about another person's mental states. For example, the sentence 'Stuart believes that she likes shortbread' describes Stuart's attitude (belief) to the proposition 'she likes shortbread'. This is completely different from the sentence 'Stuart likes shortbread', which is not an attitude to a proposition, and although it may be a good account of Stuart's disposition to eat shortbread, it is not truly an element of common-sense psychology unless someone else believes it and uses it to understand my behaviour. The form and nature of these attitudes is of fundamental importance to an adequate description of common-sense psychology.

With this grounding, I will look in more detail at two representative versions of the main competing approaches to folk psychology, Fodor's "representational theory of mind" and Dennett's "notional attitude psychology". These fall on different sides of the fence of intentionality; that is, they differ in the kind of 'aboutness' they depend on. The representational theory of mind is committed to a sentence-based description of common-sense psychology, where these sentences are mental representations with real—"intrinsic" (Searle, 1983)—intentionality. Dennett's notional attitude psychology, on the other hand, is committed to an ascribed "notional world" (Dennett, 1978) psychology where there is no such thing as real intentionality. Each view has its advantages and disadvantages.

Finally, there are a few other considerations relevant to a philosophical analysis of common-sense psychology. Firstly, there are evolutionary arguments which favour some descriptions of common-sense psychology over others—namely those which are the more biologically plausible. And second, there are methodological issues which need to be reviewed; it is important to be clear what this thesis is setting out to do—or, more precisely, what it is *not* setting out to do; that is, it is

not setting out to construct a complete causal account of human common-sense psychology. But before getting into these deeper waters, I'll start with the biggest question: what really is folk psychology?

Ontological status of common-sense psychology

Folk psychology and folk physics have much in common. They are 'mundane', that is, you don't need a degree in theoretical physics to have a good grounding in common-sense physics. And the predictions of folk psychology and scientific psychology sometimes differ radically, just as the predictions of their physical counterparts do. There is a gap between folk and scientific theories, but what is the difference between them?

According to eliminative materialists such as Churchland (1981) and Stich (1983), very little. Folk psychology is a theory in the scientific sense because, like its academic counterpart, it is used to explain, predict, and understand the behaviour of others. It can and should be judged according to the simple criterion of explanatory success, and can be dispensed with as soon as something better is available. Because folk psychology is "stagnant science" (Churchland, 1981) (a science that has been superseded in the way that alchemy was superseded by modern chemistry) it can be 'eliminated' in favour of something with more explanatory power, such as an advanced neuroscience.

According to this principle of eliminative materialism, folk psychology can and should be dispensed with, as soon as we can find something better. Although it might seem to me that I like shortbread, a more correct and complete explanation may be found by looking at the neurophysiology of my taste system. Because of the mismatch between the explanations, and the fact that the better explanation isn't that of folk psychology, perhaps the best thing to do is to throw away folk psychology and start again with neurophysiology.

The eliminative materialist's story draws an analogy between folk psychology and alchemy, arguing that alchemy was radically incorrect, so that when a more principled chemistry came along it was dispensed with. As study proceeded, there came a point when alchemy could no longer sustain the explanations needed of it, and a revolutionary shift happened to replace it with a

different science. The eliminative materialist claims that like alchemy, folk psychology is a stagnant science and sooner or later a similar revolutionary shift will happen (Churchland, 1981), probably replacing it with a neuroscientific account of human behaviour. Note that eliminative materialism proposes a revolutionary shift to a different level of explanation: the common-sense psychological and neuroscientific descriptions are at very different scales.[1] This use of levels of explanation is problematic, not least because these different levels are, perhaps, just an artifact of our perceptions as scientists; I will return to the role of intuition in levels of explanation later in chapters 8 and 14.

Another strong argument against eliminative materialism is evolutionary (e.g. Clark, 1987). According to this, common-sense psychology evolved through natural selection (Humphrey, 1976), and is therefore every bit as real as arms and legs. Of course it is possible, even today, to construct an explanation which dispenses with arms and legs in favour of a lower-level description in terms of muscles, bones, cells, and so on, but arms and legs still survive as terms in our common-sense biology. Even if neurophysiology can, in principle, provide more complete and more correct explanations, that doesn't mean that folk psychology can be dispensed with. The evolutionary claim is that folk psychology isn't a science at all, so it can't be a stagnant one! Folk psychology isn't a theory in the same sense that scientific psychology is (Clark, 1987; Searle, 1992). For example, Clark (1987) argues that folk psychology is built on a natural faculty for understanding others which has been selected through evolution, so its status is that of a "bedrock theory" rather than a scientific one. Casey (1992) reaches the same conclusion, seeing this confusion between the different kinds of theories as a category mistake.

To these technical objections, we can add the intuitive objection that denying the reality of beliefs and desires seems to be "crazy" (Searle, 1992), in the same way that it is crazy to deny that legs are real even if there are better explanations of them in terms of muscles, cells, and so on. Eliminative

[1] In practice, as McCauley (1986) argues, revolutionary shifts in science happen within levels rather than between levels. Alchemy was replaced by chemistry, not nuclear physics. This is another reason for doubting Churchland's version of eliminative materialism. Even so, McCauley is also an eliminative materialist, but of a rather different mould. He suggests that it is more likely common-sense psychology will be superseded by developments in experimental cognitive psychology. This is still vulnerable to all the other criticisms of eliminative materialism, and in particular, there is still a confusion between the two very different kinds of theory.

materialism's attempt to deny the validity of common-sense psychology is flawed by the problem that people are common-sense psychologists.

So where can we turn? Well, if you've been convinced by the eliminative materialist view, now is the time to close this thesis, get out your brain-o-meter and go down the local neuroscientific research institute, because psychology is more or less irrelevant. I think that the technical and intuitive counterarguments are the stronger, though, and that it must be admitted that there is something real about folk psychology. Of course, even if folk psychology can't be eliminated from the everyday world, its role in scientific research may still be in doubt, but Casey and Clark argue that because folk psychology is an important part of human mental behaviour, removing it from scientific study of the mind, as the eliminativists suggest, will render that science impractical, invalid, or pointless.

So even if folk psychology is the wrong kind of theory to *be* a science of the mind, it still can't be eliminated from a science of the mind, in that the study of folk psychology can and should remain part of scientific psychology (Goldman, 1993). We can take folk psychology as a natural competence to be studied using scientific principles, and now we can turn to study the elements which go to make up folk psychology—attitudes.

Attitude psychologies

The most common description of folk psychology in philosophy is as set of propositional attitudes. A propositional attitude has three degrees of freedom: the agent holding the attitude, the type of the attitude, and the proposition. An example would be 'Stuart' (the agent) 'believes that' (the attitude type) 'she likes shortbread' (the proposition).

But as Dennett (1982) points out, it is far from clear even what a proposition is. Dennett describes three different ways of seeing a proposition: first, as a syntactic sentence-like form; second, as a set of possible worlds; and third, as structures of properties and objects in the world. What these all have in common is an appeal to Frege's view of what propositions need to be able to do; and for this all three models take a proposition as a Fregean 'Thought'—having a truth value, conforming

to an intensional language, and being 'graspable' by the mind. Unfortunately, Frege was never clear what 'graspable' actually meant. To avoid premature commitment to a single view of a proposition Dennett (1987), following Churchland (1979), takes a view of propositions such that 'graspable' means that "propositions *make a difference* to a mind" (Dennett, 1987, original emphasis).

Dennett sees the diversity of ways of seeing a proposition as symptoms of an underlying problem—that Frege's three properties are actually mutually inconsistent, and in particular that graspable propositions may not have a truth value or conform to an intensional language. For example, both Putnam (1975) and Kaplan (1980) have put forward arguments that propositions can actually depend on environmental factors without the person having the propositional attitude being aware of it, so the proposition's truth value or extension can vary apparently independently of the proposition itself. Putnam (1975), for example, shows that the referent of the word 'water' may differ on different planets, on planet Earth as H_2O, and on an imaginary twin-Earth as XYZ. The actual meaning of a proposition about water, then, might depend on which planet you happen to be on.

These problems are serious. In effect, propositional attitudes aren't sufficient to describe somebody's psychology, because their behaviour can depend on environmental factors as well. As Dennett puts it, propositional attitudes can be "psychologically inert" (Dennett, 1987). This being the case, something else is needed which can play the propositional role; another kind of attitude which is strong enough to describe a person's psychology. For this, the most common strategy is to adopt sentential attitudes, replacing the proposition with something like a sentence. Fodor is perhaps the best known advocate of this strategy. Fodor's solution (Fodor, 1980) is to recommend what Putnam calls "methodological solipsism" focusing on the agent's contribution to propositional attitudes, subtracting away the context. Fodor then takes a "Realist" stance to the agent's contribution to propositional attitudes (Fodor, 1985). According to Fodor, a Realist believes that there are mental states which respect to some degree common-sense terms like 'believe' and 'intend' and which can *cause* behaviour, and that these states have semantics which can be evaluated. Realism, then, is a way of deciding what needs to be described to represent some-

body's psychology; it asserts that states corresponding to the common-sense psychological states such as 'like' and 'intend' can be held to have a valid role in psychology. If on the other hand, Realism is false, then, in effect, common-sense psychological states can make no contribution to a science of the mind.

Fodor classifies Dennett along with Churchland and the eliminative materialists as a non-Realist, because Dennett's instrumentalism implies that, strictly speaking, propositional attitudes do not exist, only that real behaviour happens to approximate them so well that their predictions work in practice. But this is the difference between Realism (in Fodor's sense) and realism. Dennett calls himself a "sort of realist" (Dennett, 1987) about the brain, while remaining instrumentalist about propositional attitudes, and agreeing with the eliminative materialists that as far as the brain is concerned, even an approximation to belief-desire psychology is false. That is, while beliefs and desires may be good for describing people, they are probably useless at a neuroscientific level. In questioning whether people really have a common-sense psychology made up of beliefs, desires, and intentions, rather than other mental states which happen to approximate them, he is far from alone (e.g. Churchland, 1981; Samet, 1993).

Unfortunately, a sentential approach to propositions also runs into problems, because different sentential attitudes can cause identical behaviour. This is the other side of the problem of propositional attitudes discussed earlier; while identical propositional attitudes can describe different mental states, different sentential attitudes can describe the same mental state. Both kinds of attitude can be less than useful as a tool for studying and comparing people's mental states. For this reason, Dennett (1982) retreats from both the propositional and sentential interpretations to find a way of capturing the similarity in different people's beliefs which deals with the implicit environmental references in propositional attitudes, and which deals with the problems of syntactic equivalence in sentential attitudes. His solution is to propose "notional attitudes" as an intermediate between (semantic) propositional attitudes and (syntactic) sentential attitudes. Notional attitudes differ from both, being ascribed by an observer living in the same environment; thus strictly they are ascribed 'as if' attitudes, but in Dennett's view the fact that the ascriber and agent share an environment and a certain amount of rationality constrains notional attitudes into validity.

Dennett's assumption of rationality has not been lost on others; this is really a soft target. In fact, Dennett's argument is rather stronger than it might seem. It is rooted in an evolutionary argument: agents—and ascribers—which weren't 'rational' would have been eliminated through natural selection. Although it is logically possible that irrational agents could exist, in practice they would have been eliminated by evolution. Unfortunately, evolution isn't always that clean, and many animals will only be rational most of the time, or only for most of the animal's behaviour, so, therefore, notional attitude psychology could only ever be accurate most of the time.

This still doesn't clear everything up. Dennett (1987) sees notional attitudes as the foundation for artificial intelligence, and explicitly uses Winograd's SHRDLU as an example, but clearly evolutionary arguments can't be applied unchanged to constructed rather than evolved artifacts, such as this. It is this freedom of ascription implied by notional attitudes that gets McCarthy (1979) into all sorts of difficulties with thermostats. Notional attitudes may be all very well, but we need some bounds on their application. On one hand, it is very clear that SHRDLU's potential grasp of propositional attitudes is extremely dubious, and on the other, the agreement between SHRDLU's world and ours seems to imply that there is something more than pure syntax involved.

So rationality in artifacts is a different matter from that in evolved creatures: in artifacts there is not necessarily any natural selection to ensure the preservation of rationality. This is an important gap in Dennett's theory. And to cap it all, it is far from clear what is meant by 'rationality' in the first place. McCarthy's definition of rationality: "it will do what it thinks will achieve its goals" (McCarthy, 1983) doesn't help much, because this is just a way of saying that belief-desire psychology works as a description of the object's behaviour. Dennett himself describes the concept as "slippery" (Dennett, 1987), and prefers to avoid defining it at all. He does, however, make one very important point linking rationality and intuition: "when one leans on our pre-theoretical concept of rationality, one relies on our shared intuitions—when they *are* shared, of course—about what makes sense" (Dennett, 1987, original emphasis). This view of rationality contrasts directly with McCarthy's, and by connecting rationality with intuition opens the door to the common-sense psychological view of rationality which I'll present in chapter 7.

Finally, Dennett's attempt to drive a wedge between the pure semantics of propositional attitudes and the pure syntax of sentential attitudes, while avoiding the restrictions of both, leads into even deeper waters. Syntax and semantics do need to be bound together, and this problem has a much wider scope than this cursory examination of different types of attitudes would imply. It is one instance of the more general problem of intentionality, so we will look at that next.

Intrinsic versus as-if intentionality

The 'graspability' of Frege's interpretation of propositional attitudes hints at a deeper problem in the philosophy of mind which surfaces here and causes all sorts of problems. This issue goes beyond the propositional attitudes, and Dennett (1987), in particular, sees it as a deep schism in the philosophy of mind. The dividing question is this simple: 'is there such a thing as original, or intrinsic, intentionality—as distinct from derived or ascribed intentionality?' One way of putting this is: 'where's the *meaning* in, say, a symbol in a computer program?' If there is a categorical distinction between original intentionality and derived intentionality, the answer would be that there could no true meaning in a program—all the meaning there is has been derived from the programmer. If there is no such distinction (which usually means there is no such thing as original or intrinsic intentionality) the meaning in a computer program is similar in kind to (although perhaps very different in degree) meaning in other people and in ourselves. This is more than a philosophical irrelevance; if nothing in a program can have true intentionality, while people do have true intentionality, there is an inevitable categorical distinction between people and programs.[2] Intentionality is the connection between an agent's propositional attitudes and that agent's environment, and any complete and correct account of propositional attitudes needs to incorporate intentionality. A solid foundation for intentionality is a necessary component of a theory of common-sense psychology.

[2] It is precisely this categorical distinction—and this distinction between original and derived intentionality—that Searle's (1980) Chinese Room argument is intended to push. I will discuss this argument in substantially more detail in chapters 5 and 11.

This is a particularly serious issue for artifacts. If we want to build an artifact capable of holding propositional attitudes, getting the intentionality into that artifact is a fundamental problem. Some, like Searle (1980), imply that this is really an effect of the physical composition and evolutionary inheritance of the artifact, so computers are simply beyond the pale. Others, like Harnad (1990), suggest that it is a balance of innate and learned components in a situated system that provides this intrinsic intentionality. Both views are subject to a number of 'slippery slope' criticisms: in fact this is the key problem with accounts based on intrinsic intentionality; movement along the slippery slope pushes both Searle and Harnad to "regress to the Cartesian vantage point" in Dennett's (1987) phrase. But Dennett's point is a good one, there really does seem to be something fundamentally dualist about the distinction between intrinsic intentionality and derived intentionality.

In fact, there are two common kinds of slippery slope argument which throw doubt on the apparently clear distinction between intrinsic intentionality and ascribed intentionality. One is the evolutionary slippery slope; if the distinction between intrinsic and ascribed intentionality is one of kind and not of degree, then if the system did not have intrinsic intentionality to begin with, either it will stay without intrinsic intentionality or there is a point where it gained it as it evolved. If there is a point where it gained it, then there was an evolutionary cause of intentionality which can be used to pin down the phenomenon, and which can, in principle, be incorporated into artifacts. The second kind of slippery slope argument is similar, but learning is responsible for the gradual change rather than evolution. In this case, there would be a point in the learning process which would be the cause of intentionality. In practice, there does not appear to be any such qualitative distinction, either in evolution or in learning—at least, not one that has anything like consensus support.

Much of the strength of the case for intrinsic intentionality comes from the case of artifacts: thermostats are artifacts, people aren't; and in contrast to a thermostat a person really does seem to have true intentionality. The case for something like intentionality in a thermostat, though plausible, is not argued strongly by many (Sloman, 1993). Even a chess-playing computer that has

been programmed to beat humans, for example, seems only to have a derived goal inherited from its programmer (Dennett, 1987), making it fundamentally different to a human chess player, which has the real goal of winning.

Dennett turns the tables on this argument with a thought experiment that points out, following Dawkins (1989) that even humans can be considered artifacts; survival machines for genetic material, so "our intentionality is derived from the intentionality of our 'selfish' genes" (Dennett, 1987). This is a somewhat evasive response, because the way we humans construct artifacts is different from the processes of natural selection and embryology: thermostats and chess playing computers don't usually evolve—at least not in the same way that animals do. It might not be the physical medium that matters in the form of intentionality, but the design process or the architecture (Sloman, 1993). There are design processes (genetic algorithms, for instance) which mimic natural selection: do their constructs qualify, or are they still just made of the wrong kind of stuff?

According to Searle (1992), yes! The first person point of view is required to achieve Fregean 'graspability', consciousness is needed for intentionality, and this requires the right kind of causal links between the brain and the world—whatever they may be. Fodor agrees on this point; the fountain of original intentionality, even if it isn't consciousness, hasn't yet been found, nor does he know where to find it (Fodor, 1981). Harnad also accepts the distinction between original and derived intentionality, but sees it more of an architectural problem than a material one (Harnad, 1990). Harnad's claim—that sensorimotor grounding can provide original intentionality—is both intuitive and persuasive, but finally suffers from the same slippery slope arguments as Searle's.

In contrast Dennett takes a relativistic view, saying that intentionality is in the eye of a beholder taking the "intentional stance" (Dennett, 1971) with an assumption of rationality. Unfortunately, the power of the intentional stance is such that it is possible to see intentionality in the most psychologically arid places. McCarthy (1979) uses intentional language about a thermostat, but without being particularly convincing: his description says more about McCarthy than it does about thermostats. Rationality as a constraining factor is not always sufficient, even for people, but for artifacts where rationality itself is doubtful, a stronger guarantee on the validity of the intentional stance is needed.

A naturalistic theory of intentionality (e.g. Millikan, 1984; 1993), on the other hand, would deny the distinction between original and derived intentionality rather differently. For Millikan (1984), intentionality is grounded in external natural relations; in the history, both the evolutionary and the social history, of the organism. Symbols and sentences, apparent bearers only of derived intentionality, are themselves part of these external natural relations, and, therefore, products of this evolutionary and social history. They, too, have a real "basic" (Millikan, 1984) intentionality. The corrolary of this is that systems which do not have an evolutionary or social history, even if they look exactly as if they ought to (for example, if they look exactly like a human) do not have intentionality at all (Millikan, 1984). This is, even for Millikan, "unintuitive". Yet, for the purposes of this thesis, there is no substantial difference between Millikan's and Dennett's positions. Both agree that there is no categorical difference between original and derived intentionality, and that evolution is central to intentionality. Their differences are most sharply emphasised by non-evolved artifacts, like computers and Millikan's non-evolved human, which Dennett suggests can have intentionality, so long as the system's behaviour warrants the intentional stance and rationality assumption.

So perhaps there is no such thing as 'original' intentionality; intentionality may come in degrees, which in a strong sense manifests itself as original intentionality and in a weaker notional sense manifests itself as derived, or ascribed, intentionality. Even so, there is a fundamental psychological, even categorical, distinction between the notional world that is applied to oneself and the notional worlds that are ascribed to others. The evidence for this is mixed. Searle's arguments are not especially convincing, because of their vulnerability to various slippery slope arguments, but there does still seem to be something very like a qualitative difference between 'real' and 'nonreal' notional worlds. Whether the distinction between original and derived intentionality is categorical, as argued by Searle (1980) and Harnad (1990), or whether it only seems to be, as argued by Dennett (1987) and Perner (1991), is still an open question. Adopting only ascribed intentionality—as Dennett does, for instance—doesn't necessarily mean that there is no finer classification of notional worlds, different gradations of status for these notional worlds; so that one's own notional world might still seem to be qualitatively different to a notional world ascribed to someone else. My money is still on the slippery slope argument and on this distinction not being truly categorical, but I'm not yet ready to bet my salary on it.

For these reasons, on the whole I am not sympathetic with the view that there is such a thing as original or intrinsic intentionality. It seems too anti-Darwinian, too arbitrary, and too dualist to be plausible. But I do suspect that there is something which closely resembles it, in that the 'quantity' of intentionality is a measure of the quality of the fit between the ascriber's mind and the system's behaviour. When the system is very like the ascriber—or when they share the same physical environment (Harnad, 1990) or history (Millikan, 1984)—the result will look very much like intrinsic intentionality. When the fit is poor, with people thinking about thermostats and chess playing computers, the result will look very much like derived intentionality. With this view, there is a whole spectrum of possibilities in between, and learning can even move in this space. This admits of both evolved organisms and artifacts having intentionality, but to radically different degrees.

Fodor's folk psychology

Fodor (1980; 1985) proposes a "representational theory of mind" for folk psychology; this is based on a sentential interpretation of propositional attitudes. Fodor's justification for this is twofold. First, it puts a theory behind propositional attitudes—a theory which allows propositions to be generated by being composed from a set of constitutive symbols or tokens. And second, when combined with the computer metaphor, the representational theory of mind allows mental processes and the causal properties of propositional attitudes to be studied as well as mental states: "computers are a solution to the problem of mediating between the causal properties of symbols and their semantic properties" (Fodor, 1985).

All this sounds ideal for an artificial intelligence project of modelling folk psychology, in that all the technological problems go away: we already know how to generate sentences from sets of constitutive symbols and how to use them causally with computers. Somewhat paradoxically, though, Fodor doesn't like artificial intelligence at all, or, rather, he feels that artificial intelligence—along with philosophy, cognitive science, and every other kind of study of the mind for that matter—has completely failed to find a way to connect languages to the world, or semantics to syntax (Fodor, 1981; Searle, 1980). The symbols have to have the right kinds of causal properties—

causal properties which are absent in traditional artificial intelligence—and we are right back at the problem of intentionality. Fodor, like Searle and Harnad, accepts the categorical distinction between intrinsic and derived intentionality.

Dennett (1987) notes Fodor's sensitivity to evolutionary slippery slope arguments on this issue. If the distinction between humans and paramecia (Fodor, 1986) is one of degree and not one of kind, then Fodor's categorical distinction between original and derived intentionality becomes dubious (Matthews, 1984). Fodor tries to get off the slippery slope by arguing that a specific competence determines whether or not there is real intentionality, the ability for the system to "'respond selectively' to nonnomic [that is, not bound by natural laws] stimulus properties" (Fodor, 1986). Unfortunately, he can't say how this competence might work (Wallis, 1992).

Another possible criticism of the representational theory of mind is that it is committed to explicit mental representation, in that mental contents "*must* be explicitly represented or the theory is simply false" (Fodor, 1985, original emphasis). Dennett makes a similar evolutionary criticism of explicit representation—explicit representations have to come about from things which aren't representations or which are implicit representations, but evolution forbids any "magic moment" of transition (Dennett, 1987).

So while the representational theory of mind seems ideal for artificial intelligence modelling of folk psychology, its main proponents also adhere to the distinction between original and derived intentionality; they accept the computer metaphor but reject the actual use of computers. And secondly, the denial of evolutionary issues in the representational theory of mind—although held in common in with much symbolic artificial intelligence—just isn't convincing.

Dennett's folk psychology

By contrast, Dennett's approach to folk psychology is through "notional attitudes" (Dennett, 1982). Dennett goes all out to construct a new attitude psychology that doesn't suffer from the shortcomings of either its propositional or sentential cousins, by rejecting methodological solipsism outright. Dennett's folk psychology is instrumentalist, meaning that he takes beliefs and

desires as truths "one must understand *with a grain of salt*" (Dennett, 1987, original emphasis); descriptions which simplify folk psychology, just like the notion of a 'centre of gravity' simplifies Newtonian physics. That is, they are *abstracta;* they aren't fully constituent elements of the physical or psychological world. The advantage of this approach, as Fodor puts it, is that "you don't have to answer hard questions about what the attitudes *are*" (Fodor, 1985, original emphasis). Fodor's principal criticism of Dennett's folk psychology is aimed at the rationality assumption; again, he raises the problem of designed artifacts: "not *everything* that's 'designed' is rational even to a degree. Bricks aren't, for example, they have the wrong kind of structure. The question what sort of structure is required for rationality does, therefore, rather suggest itself" (Fodor, 1985, original emphasis).

Methodologically, there is another problem with an instrumental approach to folk psychology. Accepting beliefs and desires as "useful fictions" (Dennett, 1987) may work well for the sake of predicting another's behaviour, but if you want to model beliefs and desires, it puts too much burden on the interpretation of that behaviour and the useful constitutive structures can be lost. Completely free use of notional attitudes means that "information is in the mind of the beholder" (Jackendoff, 1985).

I would not want to go that far. I am not advocating an "'anything goes' relativism" (Johnson, 1987) here, claiming that *all properties* are observer-relative, only that some common-sense psychological properties are. Yet the observer is central to Dennett's folk psychology. He introduces "notional worlds" as fictional worlds constructed by a beholder to describe a subject's psychological point of view. Dennett then claims that the construction of these notional worlds is a kind of phenomenology—not the phenomenology of Brentano and Husserl who used it introspectively to construct a first person notional world—but a *third person* "heterophenomenology" constructed from the outside through the "intentional stance" (Dennett, 1982; Dennett, 1991).

So the interpretation of behaviour can't *only* be in the mind of the beholder: there must be something there for the beholder to recognise as a point of view, and the beholder has to recognise it as such. The beholder must believe that to some extent she shares the same notional attitudes as the agent—making the assumption of rationality. Dennett's folk psychology moves the problems,

rather than solving them. Instead of worrying about Fregean 'graspability', we worry about the agreement between notional and real worlds; instead of worrying about the implications of methodological solipsism, we regress to worries about the beholder's folk psychology.

Formal agreement between sets of notional attitudes is problematic, but no more so than agreement between sets of propositional attitudes, while for sentential attitudes, although sentences are easily compared, beliefs themselves are not because, as Putnam (1975) argues, identical sentences can have different meanings in different environments. So Dennett retreats to informal agreement. Pointing out that sets of notional attitudes are egocentric, they can be compared by looking for similarities with respect to their different centres. This agreement is therefore not an all-or-nothing spectrum, but a continuous range of agreements, where I agree fully with a replica of myself, and (presumably) not at all with a rock.

Dennett is generally enthusiastic about artificial intelligence, and is quite happy with ascribing notional attitudes to Winograd's (1972) SHRDLU, for example, but I think there are problems here. Different beholders will construct different notional worlds for the same subject—and some may fail to construct notional worlds at all for subjects like SHRDLU. Still more worrying, there is the problem of "overestimation of cognitive prowess" (Dennett, 1985) in the Turing test. That is, sometimes the beholder seems to 'fill out' another subject's notional worlds with bits from her own—so the agreement between the two notional worlds is better than it should be. This will show up more clearly in the deeper analysis of the Turing test in chapter 5.

To return to the issue of rationality; as I've said, natural selection can't guarantee rationality for artifacts, so SHRDLU's claim to rationality is tenuous. For artificial intelligence systems like SHRDLU this needn't be a problem, because it is up to the beholder to make the assumption of rationality. But this disguises the issue rather than dealing with it; in conversational systems *both* sides need to make the rationality assumption when looking at the other, and it is the possible inability of artificial intelligence systems to make this rationality assumption in their own right that needs to be addressed—and which Dennett fails to do. Not every artifact is rational, or is able to make the rationality assumption; and so far, of course, we don't really know what the rationality assumption is. Until we do, Dennett's theory is only partially complete.

So while Dennett has gone a long way with notional attitudes, he hasn't really addressed Fodor's original criticism; he hasn't looked at the compositional structure or causal powers of notional attitudes. With Dennett's folk psychology we are constrained into a descriptive model of folk psychology, rather than an engineering approach to building it. On the other hand, the themes that Dennett adds into folk psychology, of the importance of the beholder and the importance of the agreement between notional worlds, are key elements of this thesis.

Evolutionary arguments

It is curious how infrequently philosophers have turned to the evolutionary aspects of folk psychology. Evolutionary considerations do affect folk psychology; an understanding of folk psychology requires not only an explanation of what it is, but it also requires an explanation of how it came about. Furthermore, the key tenet of Darwinian evolutionary explanations is that the difference between humans and animals is one of degree rather than one of kind, so this explanation must account for a gradual evolution of folk psychology. A monolithic folk psychology, like that proposed by Fodor (1980), is missing exactly this evolutionary account, and Fodor's acceptance of the distinction between real and ascribed intentionality appears to make it impossible even in principle.

The other problem raised by evolutionary considerations is that, for it to have evolved, folk psychology must be naturally selected, so it must confer an evolutionary advantage on its bearers. Folk psychology as a natural faculty must be 'for' something; but what is the benefit that it confers and how? I will return to this point in the next chapter—these evolutionary considerations have, with a few exceptions, been taken on board more by psychologists than by philosophers.

Methodological considerations

Philosophy does introduce some thorny methodological issues for cognitive science, and for common-sense psychology in particular. Choosing whether to accept propositional, sentential, or notional attitudes does make methodological commitments; for example, it is not obvious how to

build Fregean 'graspability' in a symbolic programming language because of the problem of intentionality—although it may be possible to *describe* agents which do 'grasp' propositional attitudes using symbolic programming languages.

Clark (1987; 1988) is especially clear on these issues. He sees the confusion between the "descriptive project" and the "engineering project" as the root flaw in both eliminativism and sentential versions of folk psychology. That is, it is perfectly valid to use the term 'belief' to describe properties of a system without requiring that the system be built from elements which *are* 'beliefs'. The eliminativists say that since beliefs aren't real at the engineering level, we shouldn't use the term descriptively; sententialists say that since beliefs are real at the descriptive level, systems must be built from them at the engineering level.

Because of this potential confusion, and because there isn't yet enough information to construct a good enough description of human common-sense psychology for an engineering project to be valid, this project will be principally descriptive. This makes it legitimate to use elements like 'beliefs' in our models, without being committed to their existence at a neurophysiological level. Although keeping the project descriptive substantially reduces its scope, if it can help to yield a good enough description, this may, in turn, assist an engineering project.

Summary: attitude psychologies revisited

Common-sense—or folk—psychology is the faculty people use to recognise one another's mental states, and to predict how their mental states will affect their behaviour. In recent years its philosophical importance has grown as we have attempted to stretch ascription of the mental to animals and to machines. This has left the scientific status of folk psychology open to question, but also shown its importance to contemporary philosophy and psychology.

I have rejected eliminative materialism on technical and intuitive grounds—although it has not been completely forgotten, and I will return to this theme in chapter 14. This makes folk psychology a legitimate subject for scientific study; but more than this, it also reveals that it is an important element of human behaviour, so not only *can* it be studied, it *should* be studied in any science of the mind.

Of the different kinds of propositional attitude, they each have their merits. 'Pure' propositional attitudes do, it must be admitted, have a big gap in the undefinability of Fregean 'graspability'. A sentential form, as proposed by Fodor, does overcome some of these problems, but at the cost of adding new ones, and in the form espoused by Fodor, without really dealing with graspability at all. A notional form, as proposed by Dennett, seems to overcome the problems of graspability, but at the cost of abandoning the constitutive structure of sentential attitudes, which is the main advantage of the sentential form.

The obvious solution is to try to combine the benefits of the two, but Dennett doubts the actual constitutive structure proposed by Fodor's representational theory of mind, sometimes quite strongly, seeing attitudes as "a useful—if sometimes treacherous—approximation that is systematically incapable of being rendered precise" (Dennett, 1987). This, it must be admitted, is a real problem—if the nexus of beliefs, desires, and intentions is not the right constitutive structure then what else should we be looking for? One way of seriously addressing this is to look at the attempts to describe real folk psychological behaviour, and to see to what extent it does fit in with beliefs, desires, and intentions. Two aspects of this approach, in psychology and artificial intelligence, will be analysed more fully in chapters 3 and 4 respectively.

Even ignoring the differing opinions on the actual ontological status of beliefs and desires, combining the advantages of the sentential and notional versions of folk psychology is not trivial, because evolutionary considerations rule out some aspects of Fodor's theory, and Dennett's theory rests on some assumptions that are questionable for artificial (non-evolved) systems. A general theory requires a more unified approach to natural and artificial systems, and this has a number of implications, both evolutionary and methodological.

Before passing on to these analyses, there are a number of methodological issues which need to be recognised. Evolution is one: Fodor's representational theory of mind isn't really backed up substantially by evolutionary explanations, where Dennett's "notional attitude psychology" does accept evolutionary considerations—this is another reason for questioning the ontology of Fodor's representational theory of mind. Psychologists are generally more concerned about evolutionary explanations than philosophers, so this point will be addressed more fully in the next chapter. A second, more general, methodological issue is of the kind of study these characterisations of folk psychology permit through artificial intelligence techniques. Ultimately, both Fodor's and Dennett's positions imply the rejection of all but descriptive approaches, but for very different reasons. I accept these points, and intend this thesis to be firmly within the descriptive framework, but in the spirit of artificial intelligence, I believe that even in a descriptive project some design considerations should be taken into account.

This philosophical groundwork highlights many of the definitive characteristics of common-sense psychology, but it also shows many of the gaps and inconsistencies in the different theoretical and methodological approaches. One of the best ways to highlight and clarify these is to look at how people use common-sense psychology in practice, on an everyday level. Research on this has become an equally important issue in another discipline, psychology, so I will turn to that in the next chapter.

Chapter 3

Common-sense psychology in psychology

Introduction

The review of the philosophy of common-sense psychology in the previous chapter doesn't answer many of the questions that need to be answered. Although it provides some clues to the nature of common-sense psychology—and hopefully warns us against making any serious category mistakes—it doesn't really provide any details of the actual behaviour of a common-sense psychology; *how* it does what it does. While it is all very well to discuss what might underlie common-sense psychology in principle, there is no substitute for testing hypotheses about how common-sense psychology works in practice. Without this line of research, there is a real danger that our understanding of common-sense psychology may be ungrounded in real competences. To help us answer questions on this complementary path, we must turn to a different discipline: psychology.

Psychological research in common-sense psychology has followed two main themes, in comparative psychology and developmental psychology, and in both the scale of social reasoning has revealed much of what we tend to take for granted in adult humans. In both areas, research has followed the more traditional psychological plan, making hypotheses and then testing them experimentally; and both areas have, in general, taken evolutionary plausibility to be an important criterion when evaluating competing hypotheses.

Many of the differences between philosophical and psychological research have been terminological, where philosophy talks about folk psychology, psychology talks about theory of mind; where philosophy talks about intentionality, psychology talks about representation. But these terminological differences reveal differences in emphasis; psychology is less interested in the prob-

lem of intentionality, of *aboutness*, in propositions, but more interested in the mechanics of representations. And some of the more subtle differences in philosophy, for example between sentential and propositional attitudes, are largely ignored in psychology. Psychology, in its turn, adds new themes, such as a significant emphasis on egocentricity; that is, whether people's common-sense psychology develops by extending 'first person' ascription to other people, or whether it develops by extending 'third person' ascription to oneself.

To begin this look at the psychology of common-sense psychology, I will first turn to research in comparative psychology; because it was the evolutionary perspective that comparative psychology offered which perhaps revealed the magnitude of common-sense psychology for the first time.

Common-sense psychology in animals

The study of common-sense psychology in animals was first legitimated by Darwin's argument from natural selection that "there is no fundamental difference between man and the higher animals in their mental faculties" (Darwin, 1871). Darwin goes through a remarkable list of apparently human mental powers, ranging from language to self-consciousness, in all cases claiming that the difference is "in degree rather than in kind" (Darwin, 1871).

Darwin's (1872) study of the expression of emotions builds on this framework. Darwin accepted the dominant psychological view of his day that the gestures which express emotions were derived from ancestral equivalents, and were retained for the purpose of expressing emotions. This view persisted until the turn of the century, both in human and in animal psychology, which drifted to become blatantly anthropomorphic (e.g. Romanes, 1886) and which consistently and often inappropriately took animal mental states to be the same as human ones (Krementsov & Todes, 1991). In the limit the reaction against this led to behaviourism.

During most of this century, however, the behaviourist domination of psychology prevented most comparison between human and animal psychology—indeed, it effectively ruled out any study of common-sense psychology at all through its denial of the use of mentalistic terms even for hu-

mans. The advent of the computer metaphor eventually made the use of mentalistic terms such as 'memory' legitimate again and lead to the overturning of the behaviourist domination of psychology (Miller, 1983; Turkle, 1988).

Building on the Darwinian foundation Humphrey (1976) pushed common-sense psychology into a new prominence. Under the label "natural psychology", his theory was that much of human—and higher primate—intelligence is directed to the understanding and manipulation of others in a social group. After all, even for early humans, dealing with the physical environment hardly seems to merit all the complexity of human mental powers.

The roots of Humphrey's theory go back a long way in social psychology and ethology. Mead, Chance, Jolly, and Kummer have all—and apparently independently—reached similar conclusions. This collective view is described by Byrne and Whiten (1988) as "Machiavellian intelligence". The idea is simply that animals which are better at predicting and manipulating each other's mental states—and therefore indirectly manipulating each other's behaviour—will gain an evolutionary advantage over those other animals, and will be favoured by natural selection. The result is an evolutionary "arms race" (Dawkins, 1989) leading to the complexity of human common-sense psychology and human society. The Machiavellian intelligence hypothesis has been target for some criticism (e.g. Ridley, 1993), notably in that the 'arms race' argument should also apply to other great apes, and therefore doesn't provide a complete account of the difference between human and other kinds of intelligence. But this controversy is perhaps outside the scope of this thesis—all that matters for my purposes is that humans have a common-sense psychology and the Machiavellian intelligence hypothesis supports this. The counterarguments are aimed at the grander claims sometimes made on the same hypothesis, namely that *all* the mental differences between humans and other animals can be accounted for by the Machiavellian intelligence hypothesis.

The connection with the study of common-sense psychology in humans was made by Premack and Woodruff's (1978) experiments to determine whether chimpanzees have a common-sense psychology "theory of mind". Premack and Woodruff argued that their chimpanzee subject, Sarah, could predict and explain a human's actions in terms of mental states such as intentions,

and use this understanding to deceive a hostile trainer. Their experiments were criticised methodologically by Bennett (1978), Dennett (1978), and Harman (1978), among others, who all claimed that an understanding of 'false beliefs' was a necessary indicator for a true theory of mind—in other words, they needed to prove chimpanzees capable of ascribing a mental state which is different from their own mental state. Without this, Sarah has not been proven to have *intentionally* deceived, only to have behaved so as to achieve a deceptive effect—behaviour which could have been learned without any understanding of other minds. This proposal lead to the first test of false belief in humans (Wimmer and Perner, 1983), and enabled the transition to developmental psychology in the study of the theory of mind.

Since Premack and Woodruff's challenge, many studies of different animals' possible theory of mind have been made (e.g. Byrne & Whiten, 1988; Cheney & Seyfarth, 1990; Cheney & Seyfarth, 1991; Gómez, 1991; Ristau, 1991). None of these are conclusive. Cheney and Seyfarth, for instance, say "some of our results argue against a 'theory of mind' in monkeys, while others support it, and still others are inconclusive" (Cheney & Seyfarth, 1991). The problem is that many aspects of common-sense psychology are apparently common in animals. Deception, for example, is one possible aspect of common-sense psychology which can be found in many animals (Byrne & Whiten, 1988; Ristau, 1991), but which can often be explained more simply by learned association. Other possible signs of common-sense psychology—imitation and pretence, for instance—are open to similar methodological problems (Whiten & Byrne, 1991).

The evidence is against any non-human animals having a theory of mind on the human scale, in that individuals of no species seem to have the complete capacity for reasoning about and acting on their conspecifics' beliefs, desires, and intentions. Even so, there is evidence that some animals, and particularly chimpanzees, have substantial components of a theory of mind (Whiten & Byrne, 1991), although even they have more difficulty ascribing informational states (such as beliefs) than they do motivational ones (such as desires) (Premack & Dasser, 1991). But despite the methodological problems of studying common-sense psychology in animals, there are at least two contributions it can make to an understanding of common-sense psychology in humans; first, the comparative perspective can offer insight into its possible evolutionary origin in humans (Gómez,

Sarriá, & Tamarit, 1993); and second, research on animals is a rich source of theoretical and methodological ideas for studying it in humans. Comparative psychology, therefore, may complement, but is no substitute for, research on common-sense psychology in humans.

Common-sense psychology in humans

Common-sense psychology does not appear to be completely innate, or at least, if it is, it goes through several stages of development. There are fundamental differences between an adult's ascription of a mental states and a young child's. Young children of about two and a half show an apparent egocentricity; that is, a young child seems unable to distinguish between their own point of view and anybody else's, and behaves as if they use their own knowledge of their own mental states to reason about other people's behaviour, projecting their own mental states onto other people.

According to this egocentric view—which is Cartesian in style (Perner, 1991)—we have direct access to our own minds, and the terms we use for mental states refer to internally known states. With time, we get better at using these terms for others, by taking their point of view; we use our direct knowledge of our own minds to predict the mental behaviour of others. This role-taking is intuitive, but it leads to the philosophical 'other minds' problem; since other people's minds are opaque, how can we know that the same terms are applicable? That is, how can we know that other people construe the world the same way we do?

But there is an alternative view, in which terms for mental states are part of a coherent 'theory' which we use for predicting one other's behaviour. There is no egocentric extrapolation from the first person perspective to the third person perspective, instead, the first person and third person perspectives develop hand in hand. By and large, this 'theory theory' behaves almost identically to the egocentric view; behavioural distinctions between them are very subtle (Perner, 1994), and devising experiments to unambiguously distinguish between them can be very hard.

Buried deep in this distinction there is a point of fundamental importance. Which has primacy—the first person point of view or the third person point of view? Do we discover others through ourselves, or ourselves through others? That is, as children, do we learn how we ourselves think by imitating and learning from others, or do we learn how other people think by extrapolating from ourselves. In their extreme forms neither of these alternatives is completely satisfactory; ultimately, the purely egocentric first person stance leads to dualist egoism, and equally ultimately, the purely allocentric third person stance leads to rampant relativism.

To help answer these questions, developmental psychology has mainly looked at the stages through which a child's common-sense psychology seems to pass as it develops and matures. There are several clear patterns which emerge. First, as Wellman (1991) shows, there is a distinct change in children's language between the ages of two and a half and four. Secondly, a 'false belief' test—derived from Bennett's, Dennett's, and Harman's criticisms of Premack and Woodruff's experiment—also shows a clear change in children's ascription of mental states to others over roughly the same time. Finally, the developmental disorder of autism shows characteristics of a specific deficit in common-sense psychology, so it too has been thoroughly investigated for its effects.

Beliefs, desires, and intentions: belief-desire psychology in psychology

The dominant approach to the study of common-sense psychology in psychology, as in philosophy, is through categorising human mental states as a set of beliefs, desires, and intention with respect to actions. The connections between them are shown in figure 3.1. A belief-desire psychology like this is not intended to be a scientific description of how people behave—that is, it is not part of a scientific psychology. It is a model of how a common-sense psychology construes other people's mental states—not of how they actually are. While belief-desire psychology is a useful way of dividing and understanding human mental states, as shown in figure 3.1, it also shows how those mental states are connected to the world, and to other apparently neurophysiological states such as hunger. For a more complete description of the belief-desire model, see Wellman (1990).

Belief-desire psychology seems neither to be constant nor innate (for all that Fodor, 1985, claims). Figure 3.2, after Wellman (1991), shows the dramatic increase in children's natural language use of belief terms (e.g. 'think') between the ages of two and a half to four, while the use of desire terms (e.g. 'want') is present from a much earlier age. Individual children follow a similar pattern (Wellman,

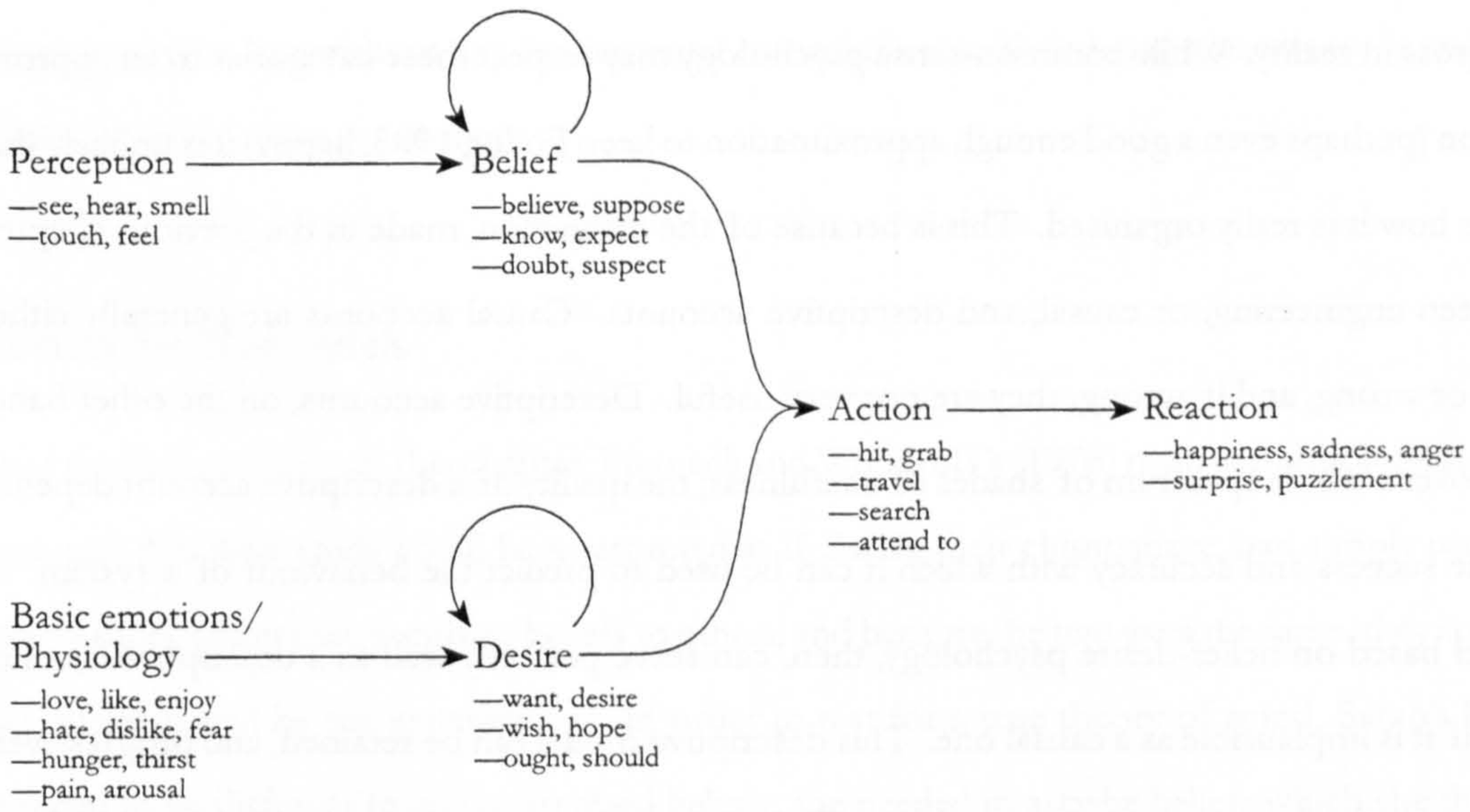

Figure 3.1. Beliefs, desires, and actions (after Wellman, 1990)

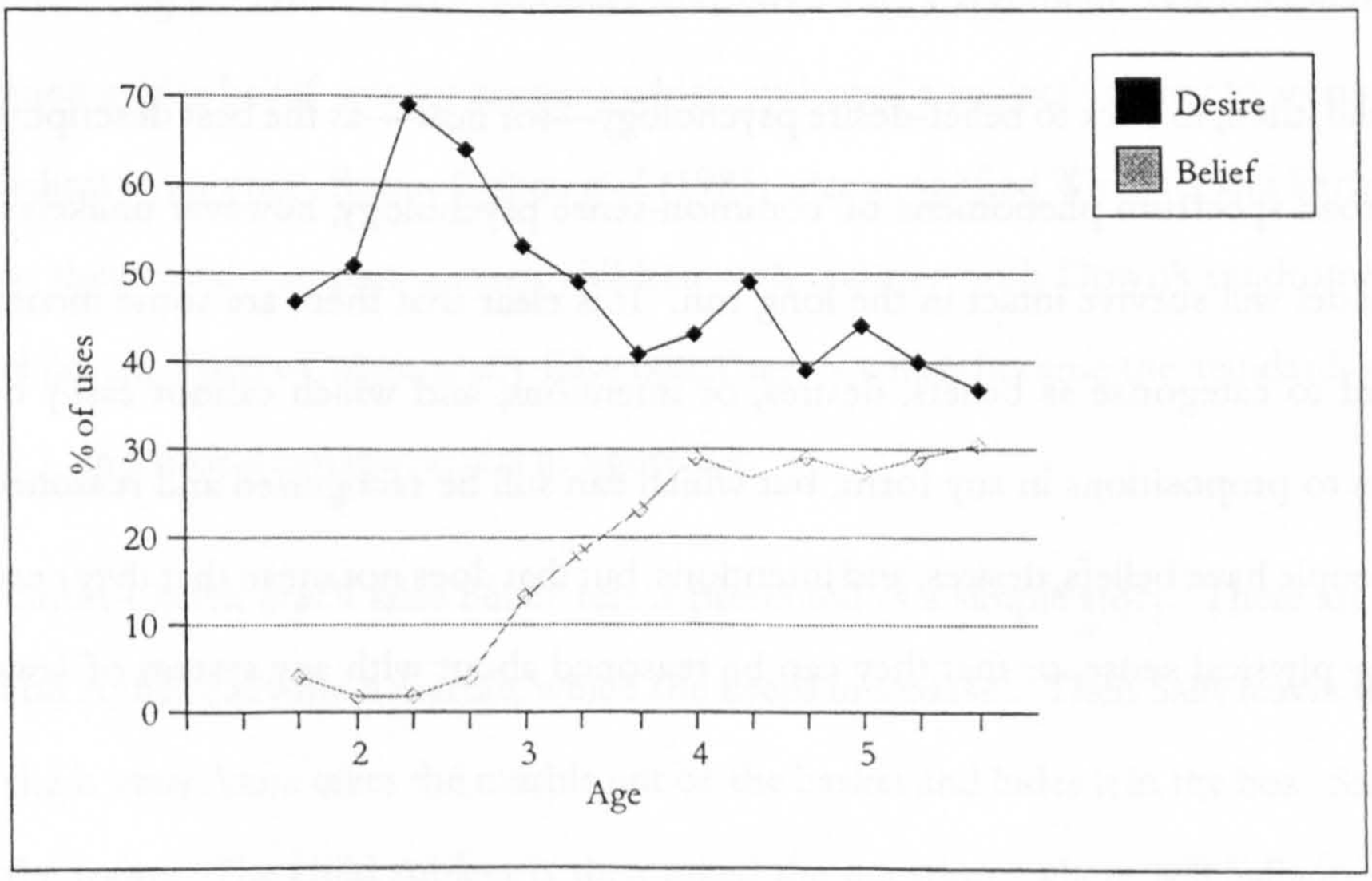

Figure 3.2. Changes in belief-desire reasoning (after Wellman, 1991)

1991). So, although there is a fundamental change, there is no evidence for any "graduation day" (Bruner & Feldman, 1993) shift from a "simple desire psychology" (Wellman, 1991) to a belief-desire psychology.

To categorise all human mental states as beliefs, desires, or intentions seems rather ontologically violent. It is. But that isn't what belief-desire psychology does. It merely claims that our common-sense psychology works in terms of members of these three categories. Nevertheless, along with Dennett (1987), Hobson (1993), and Samet (1993) I think that this categorisation is probably too gross in reality. While common-sense psychology may respect these categories to an approximation (perhaps even a good enough approximation to keep Fodor, 1985, happy) it is unlikely that this is how it is really organised. This is because of the distinction, made in the previous chapter, between engineering, or causal, and descriptive accounts. Causal accounts are generally either right or wrong, and if wrong, they are not very useful. Descriptive accounts, on the other hand, can cover a whole spectrum of shades of usefulness; the quality of a descriptive account depends on the success and accuracy with which it can be used to predict the behaviour of a system. A model based on belief-desire psychology, then, can serve perfectly well as a descriptive account even if it is implausible as a causal one. This descriptive model can be retained, and progressively revised until a better account is available. This better account might, or might not, involve a paradigm shift to a completely new causal framework, or it might just involve a redescription within the existing framework

It is completely valid, then, to stick to belief-desire psychology—for now—as the best descriptive account for the broad spectrum phenomena of common-sense psychology, however unlikely it seems that this model will survive intact in the long run. It is clear that there are some mental states that are hard to categorise as beliefs, desires, or intentions, and which cannot easily be framed as attitudes to propositions in any form, but which can still be recognised and reasoned with. Of course people have beliefs, desires, and intentions, but that does not mean that they need to be 'states' in any physical sense, or that they can be reasoned about with any system of laws.

Moods are one of the best examples of these; moods are essentially dispositional rather than categorical — tendencies to particular behaviour patterns rather than behaviour patterns in their own right (Ryle, 1949). A combination with a dispositional model is, therefore, probably inevitable.

So while there may be something dubious about the idea that common-sense psychology should be thought of only in terms of beliefs, desires, and intentions, we will not abandon this model completely. The results in figure 3.2 are striking enough to show that it is a pretty good approximation. So, for the sake of this project, I will adopt beliefs, desires, and intentions as a good approximation, but without any fundamental commitment to their being a necessary foundation for a common-sense psychology.

Testing for false beliefs

As I mentioned earlier in this chapter, Premack and Woodruff's (1978) main methodological problem was that their study could be interpreted as if Sarah, their chimpanzee, had simply used her own beliefs, rather than ascribing beliefs to others, and because the two were the same, they couldn't be distinguished by the experiments. In order to test for a true theory of mind, Sarah's beliefs needed to be different from the ascribed beliefs; she needed to ascribe beliefs which she thought were false.

Following the criticisms of Premack and Woodruff's argument, Wimmer and Perner (1983) devised a false belief test for humans which evaluated a subject's ability to ascribe definite but false beliefs to another. Baron-Cohen *et al.* (1985) later simplified Wimmer and Perner's false belief test so they could compare autistic children with children with Down's syndrome, and with normal children. Baron-Cohen *et al.*'s false belief test—which became the standard—is shown in figure 3.3, and their dramatic results in figure 3.4.

Baron-Cohen *et al.*'s false belief test is presented as a simple story. There are two puppets, Sally and Anne. Sally has a marble, which she keeps in a basket. Then Sally leaves the room, and while she is away Anne takes the marble out of the basket and hides it in the box. Sally comes back into the room. The child subject is then asked the question "where will Sally look for her marble?"

Older children will say that she will look in the basket, because although they know that the marble is now in the box, they also know that Sally doesn't know that it has been moved from the basket, and they are capable of distinguishing between the two. Younger children, aged three, for example, say that Sally will look in the box because they do not ascribe the false belief to Sally.

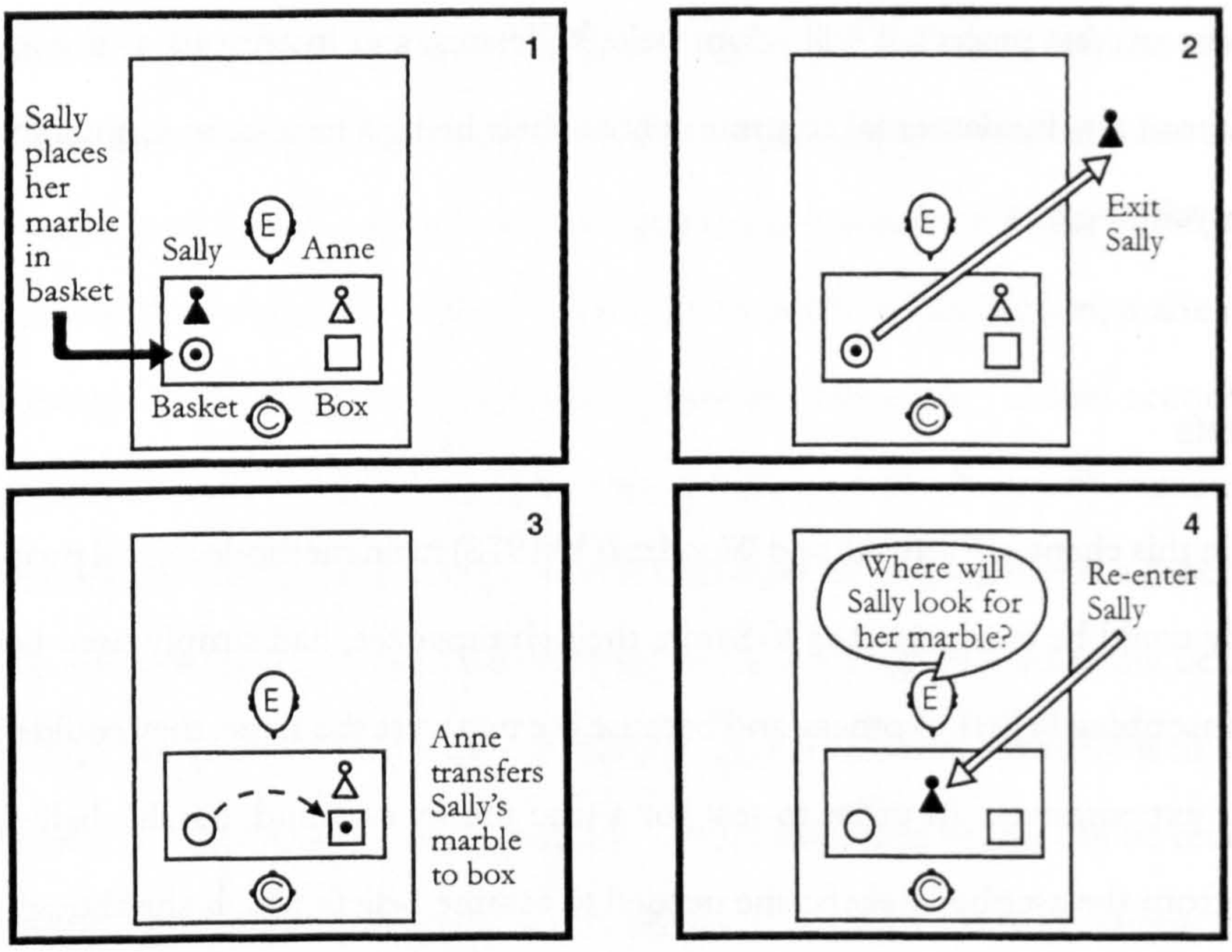

Figure 3.3. Baron-Cohen *et al.*'s false belief test (after Baron-Cohen *et al.*, 1985)

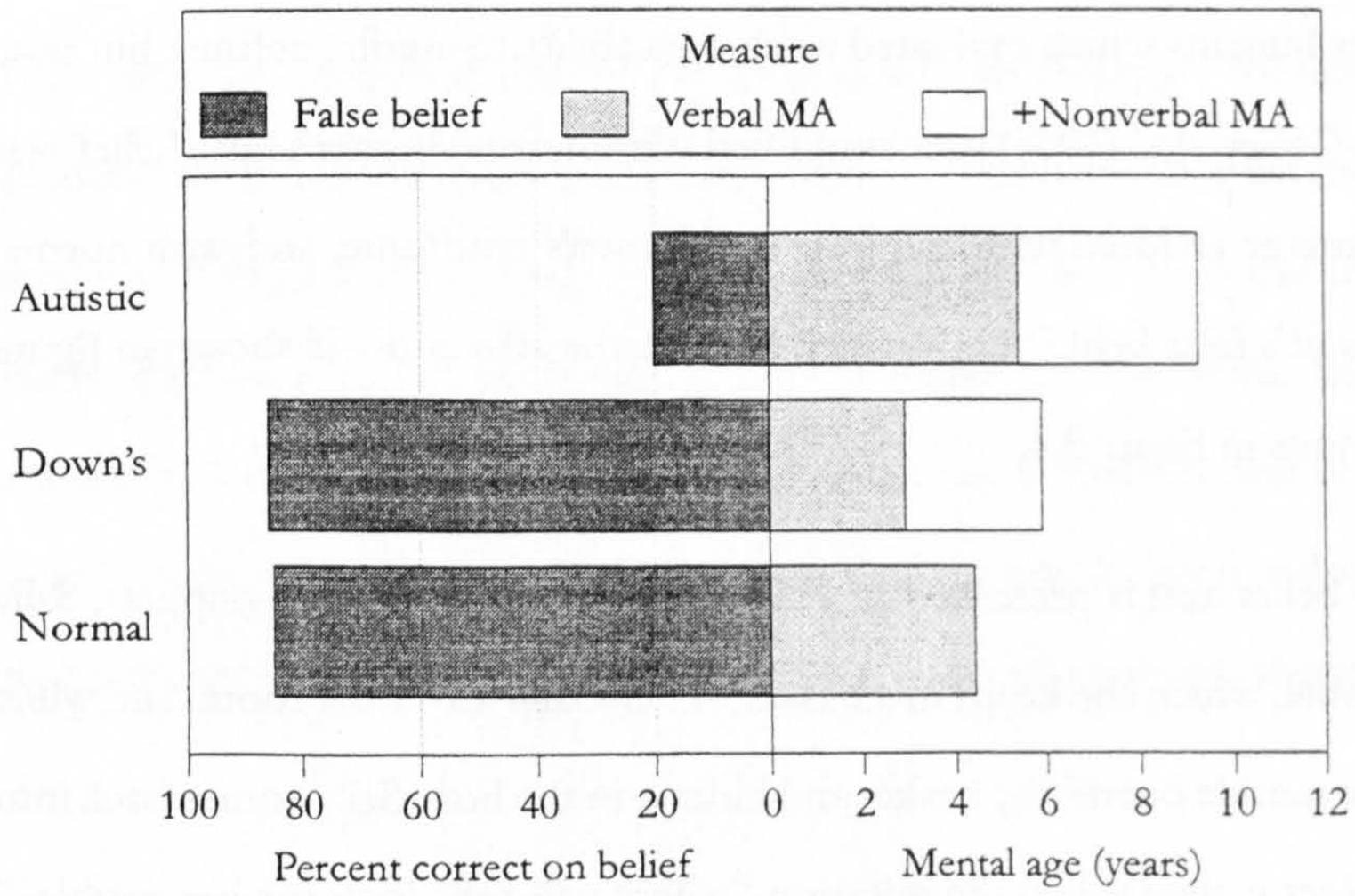

Figure 3.4. Results of the false belief test (after Perner, 1991, data from Baron-Cohen *et al.*, 1985).

Figure 3.4 shows that both normal children and those with Down's syndrome can achieve over 80% success on the false belief test, but autistic children—even with a greater verbal mental age—only pass the false belief test 20% of the time. Autism's specificity to this test was further confirmed when those autistic children who passed this false belief test were shown to fail a more complex false belief test involving embedded beliefs (e.g. 'John believes that Sally believes that...') (Baron-Cohen, 1989).

There are other kinds of false belief tests. For example, Astington and Gopnik used a similar approach to study children's understanding of their own previous mental states compared to their understanding of other people's current mental states. It turns out that a very similar kind of development happens, and the patterns of ascription to oneself follow similar patterns to the ascription to others (Astington & Gopnik, 1991). Astington and Gopnik believe that this argues against the plausibility of the purer versions of both the theory theory (Leslie, 1991) and the simulation theory (Harris, Brown, Marriot, Whittall, & Harmer, 1991; Johnson, 1988). Instead, they suggest that these results imply that the first person and third person ascription is developed simultaneously (Astington & Gopnik, 1991; Perner, 1991; Wellman, 1991) and use the same underlying mechanism. These models neatly escape the "regress to the Cartesian vantage point" (Dennett, 1987) which leaves unexplained the connection between the first person and third person perspectives.

Certainly something significant happens to these children between the ages of two and a half and four, although precisely what is as yet unclear. A number of other points need to be made. First, some kinds of mental impairment (e.g. Down's Syndrome) make no apparent difference to the false belief test, age for age, at all. On the other hand, autistic children remain unable to ascribe false beliefs, even when much older then four. All this seems to show that this one development disorder—autism—might indeed have a significant connection with human common-sense psychology.

Autism

If common-sense psychology is to any extent innate and modular, then it would seem entirely probable that there could be a specific disorder in this area. Autism is a likely candidate: it is diagnosed by difficulties in coping with social interaction and both verbal and nonverbal communication. Autistic children are as good as normal children when solving problems involving common-sense physics (Baron-Cohen, Leslie, & Frith, 1986) and almost as good with the ascription of mental states not requiring a theory of mind (Perner, 1991). All this seems to show that the impairment is specific to common-sense psychology. Against this theoretical background, the many experiments that have been carried out comparing autistic and normal children have proved a rich source of models and architectures for common-sense psychology (e.g. Leslie, 1991; Perner, 1993).

According to Baron-Cohen *et al.* (1985) and Leslie and Roth (1993), autistic children have a specific deficit in the use of 'metarepresentation', by which they mean mental propositional attitudes. Leslie and Roth (1993) suggest that there are three possible modes of impairment: in the development of the attitudes, in the ability to form propositions into propositional attitudes, and in the ability to use propositional attitudes; any of these could cause many of the apparent symptoms of autism.

Perner (1993) takes a similar stance, again suggesting that autism is a deficit in the use of metarepresentation—but this time in Pylyshyn's (1978) general sense as a representation of a representation. This contrasts with Leslie and Roth's interpretation of the term 'metarepresentation' as something specifically mental, deployed by a theory of mind. Leslie and Roth take children to draw a fundamental distinction between agents and physical objects, a distinction which Perner resists except in so far as it arises from the ascription of representations.

One of the stranger aspects of autism is that autistic children actually show pretty good understanding of false photographs, comparable to that of normal children (Leslie & Thiass, 1992; Zaitchik, 1990). Leslie's explanation of this is that all children draw this fundamental distinction between agents and objects—where only agents are capable of having propositional attitudes—so a photograph (which isn't an agent) has a different kind of representation to a person. Perner

(1993) disagrees, instead claiming that autistic children have a domain-general deficit in metarepresentation, and that the photograph tasks can be solved without metarepresentation and with only a more basic "situation theory" that all children possess and which is unaffected by autism.

But others are critical of the suggestion that autism can be explained simply by a deficit in a theory of mind. Samet for example, accepts the belief-desire-intention nexus as a cognitive science approach to the theory of mind, but with respect to autism comments "my sense is only that there is more to the disease than this" (Samet, 1993). Hobson argues that "a person's understanding of the minds of other people and herself is grounded in forms of interpersonal relatedness that are 'affective' in nature" (Hobson, 1993; Humphrey, 1984). Autism can be seen as an affective disorder as well as a theory of mind disorder. It can even be seen as a deficit in the ability to organise narrative structures (Bruner & Feldman, 1993), which might manifest itself in either of the other forms—this would, of course, restrict autism to animals capable of narration, but there is no evidence on this either way.

Autism is confusing! Most of the evidence can be interpreted in several different ways. And individual differences among autistic children are often hidden by these studies; in practice there is substantial variation between individuals, and social behaviour even among autistic children is still remarkably complex (Lord, 1993). Even though the effects of autism seem quite specific, they are specific in areas which are so fundamental to human behaviour that they are still incredibly difficult to pin down. But it does seem that whichever way autism is finally explained, it will provide important clues to human common-sense psychology.

Theory theory or simulation theory

Most psychologists describe common-sense psychology as a 'theory' in the sense that it is used to explain, predict, and understand the behaviour of others—this is why the term 'theory of mind' is used so widely in this field. But there is an alternative view which claims that instead of using a theory to predict another's behaviour, we effectively simulate another person's mental processes

to achieve this prediction. These two alternatives theories are usually called the 'theory theory' and the 'simulation theory' respectively. If we are to build models of common-sense psychology, it is important that we know which of these is most appropriate.

The theory theory (Astington & Gopnik, 1991; Baron-Cohen, 1993; Perner, 1991; Wellman, 1990) describes this ability to predict and explain another's mental states as a theory analogous to—but not the same as—a scientific theory, for example, like the way that Newtonian physics predicts and explains the motion of objects by a set of rules. The theory theory, then, is a bit like heuristic reasoning in artificial intelligence.

According to the simulation theory, on the other hand, one person's prediction and explanation of another's mental states is through an ability to simulate being the other person. According to this model, one person 'puts themselves in another's shoes' (Astington & Gopnik, 1991; Perner, 1991) to decide what they would see and know in that situation. The strongest adherents of the simulation theory are Gordon (1986), Goldman (1993), and Harris (1993). The simulation theory is close to look ahead search in artificial intelligence.

Current research is particularly active in trying to clarify the nature of the distinction between the simulation theory and the theory theory—especially where empirical studies are concerned. Actually choosing one side or the other is still fraught with danger; the jury is still out on these matters. Despite these differences, most agree on certain shared issues. First, simulation does seem to be an accurate description of certain kinds of prediction; emotions, for example (Harris, 1989; Perner, 1991). Second, there are a few cases where the theory theory seems to be a better match to experimental evidence than the simulation theory (Perner, 1994). Third, in many cases—although not all—the difference between the simulation theory and the theory theory could well be illusory rather than real (Davies, 1994; Perner, 1994). It seems, therefore, that both the simulation and theory theories may be required.

Perner (1994) makes this suggestion explicitly, but it does resonate with the use of both heuristics and look ahead in artificial intelligence's approach to game playing. By combining heuristics (a kind of theory) and look ahead (a kind of simulation), it seems likely that artificial intelligence, out

of necessity for its interest in solving 'hard' problems, may have stumbled on the same need to combine these techniques that evolution had, millennia earlier. If this is the case, it means that there are, once again, ways that artificial intelligence can contribute to the debates on common-sense psychology by helping to explore and clarify this connection.

Another problem that arises is that while the experimental evidence is very clear that there is a change in children as they mature—and that this change isn't displayed in the same way by children with autism—the actual development path followed isn't clear at all. Meltzoff and Gopnik (1993) distinguish two positions on this: "modularity nativism" and "starting-state nativism", depending on whether the development is constrained architecturally or representationally (Elman *et al.*, 1996). According to the modularity view, children's common-sense psychology is principally structured as a set of modules, and it is this set of modules that is impaired in autism; that is, children with autism are simply not mentally equipped to do ascription of mental states. By the starting-state view, on the other hand, children have special knowledge about their relation to other people, and it is this knowledge that is impaired in autism—the effect is to corrupt the evidence available to children about the connection between their own and other people's mental states, and through this to damage the ascription of mental states indirectly. Leslie's (1991) accounts falls into the first category: according to this view autism is a deficiency in a "theory of mind module". Meltzoff and Gopnik, following Hobson, take the other position and argue that autism is a deficiency in the ability to see others as persons rather than objects, and that this is caused by a gap in the evidence children have available to them as they build their theory of mind (Meltzoff & Gopnik, 1993). This is an especially important point because, as I will discuss later in chapter 7, people's ability to tell when common-sense psychology is useful—in other words, when something is acting as an agent rather than an object—seems to be an important, if somewhat neglected, area of research in cognitive science.

Perner argues that in practice both simulation and theory forms of common-sense psychology are probably required, but for dealing with different situations (Perner, 1991), and he has empirical evidence for this (Perner, 1994). Of course, it is always possible to claim that a theory can generate

simulation (although not *vice versa*) as Davies (1994) and Heal (1994) do. But even this seems unsatisfactory; it shouldn't be a question of whether a theory *can* generate simulation in principle, but whether a theory *does* generate simulation in practice.

As I've hinted, in artificial intelligence a similar distinction can be found. The simulation theory is closely related to the traditional artificial intelligence technique of look ahead search in game playing. A program looks ahead by pretending to *be* an opposing player, and using the best counter move for the opposing player in the analysis of a possible move. The theory theory, on the other hand, is more like a set of good heuristics. Cast in artificial intelligence terms, then, Perner's suggestion is analogous to the argument that successful artificial intelligence programs for playing games of any complexity need to use a subtle mixture of simulation (look ahead) and theory (heuristics).

This possible analogy between artificial intelligence's approach to game playing and common-sense psychology is a bit of a 'blue sky' hypothesis, in that there is little hard evidence for it—and given the controversy, it is a hypothesis which deserves sound evidence. Even so, the similarity is striking. In practice, simulation on its own is impractical, as without some heuristic component there is nothing which guides a system about what to simulate. Grounding this in the context of common-sense psychology, there must be a theoretical element at least to the extent that people know who to simulate and when. I will suggest, in chapter 7, that anthropomorphism might provide some of this theoretical component, but it seems likely, as Perner (1994) suggests, that there are other elements besides. This gives us more hints about the possible structure of human common-sense psychology, but it still leaves open the problem of when and how this subtle mixture of simulation and theory is created.

Evolutionary origins of common-sense psychology

There is a considerable amount of evidence for a naturally evolved common-sense psychology, even in a few animals. Although many animals seem to be able to play limited games of "guessing thoughts" (Wittgenstein, 1953)—deploying sophisticated deception, for example—in practice, only chimpanzees seem to have common-sense psychology on anything like a human scale (Whiten

& Byrne, 1991). Nevertheless, coupled with the theoretical argument that such a common-sense psychology would be evolutionarily adaptive (Humphrey, 1976) this does seem to offer strong evidence for an evolutionary origin to common-sense psychology.

The evolutionary evidence, however, is rather unhelpful when it comes to providing any indication of when and how common-sense psychology actually evolved. Most agree that higher primates seem to have a common-sense psychology nearly as powerful as that of humans, although there still appear to be qualitative differences between the two. Research outside primates, however, has been much rarer (but not absent, e.g. Ristau, 1991). According to one criterion (expression of pain being 'for' something, Humphrey, 1984) most other mammals have something like a common-sense psychology, and it may even be shared by reptiles and birds, too, although most other vertebrates and all invertebrates do not seem to qualify.

This lack of research may hint at a hidden but significant methodological issue: research complexities seem to grow as animal subjects become genetically more dissimilar to their human experimenters. This cannot necessarily be just put down to a general difference in 'intelligence'—this amounts to brushing a large problem under a small carpet. Furthermore, an appeal to a general 'evolutionary ladder' looks very much like anthropocentricity. I think there are far more subtle and important issues involved here, and I will return to this theme in chapters 7 and 13.

Developmental origins of common-sense psychology

Developmental psychology has been especially important in highlighting the possible development path of a full common-sense psychology. Agreement on these issues is only partial, but the different stages and modules leave important clues on the nature and structure of common-sense psychology. Almost nobody, however, agrees with Fodor's (1987) apparent view that common-sense psychology is down to a set of special-purpose modules present from birth. The disagreements which do exist concern the actual development process which results in adult human common-sense psychology.

Perner (1991) argues that children begin with an innate disposition to attend to the expression of mental states, which is followed by their developing first a concept of mental states as relations to situations (a situation theory), and then finally a concept of mental states as representations (a representation theory). In fact, in Perner's model, there is no distinction between the physical and mental, and therefore there is strictly no such thing as a common-sense psychology as opposed to a common-sense physics. It is merely that in practice, to make the kinds of predictions needed of a common-sense psychology requires metarepresentational faculties that aren't much needed by other aspects of human common-sense reasoning.

Wellman, on the other hand, suggests that children really are psychologists, and that they build on a simple desire psychology by learning about beliefs to deal with the anomalies that arise in a psychological model that does not account for beliefs (Wellman, 1991). Initially, therefore, young children are neither mentalists nor behaviourists: they do postulate internal states, but only desires, not the whole range of inner mental states that adults can cope with.

Baron-Cohen (1993) proposes that there may also be an attention-goal psychology as a precursor to Wellman's simple desire psychology. This would be built on an innate sensitivity to other people's attention, including gaze monitoring and an awareness that people's actions are usually goal directed. This attention-goal psychology, then, provides a sensitivity to goals which can in turn build Wellman's simple desire psychology.

But the role of attention in common-sense psychology is complex. Although it is a key part of Baron-Cohen's and Perner's models, attention is itself capable of several different interpretations. In the first place, it is both physical and mental: it constitutes both turning the senses towards something and turning the mind towards it (Gómez *et al.*, 1993). And furthermore, attention can all too easily act as an all-powerful genie homunculus in a model (see chapter 14, and Akins, 1993)—that is, it is psychologically very easy to hand wave all sorts of apparently intractable psychological processes into a box, then to label it 'attention' and forget about studying it until next Thursday. This problem, which I discuss in more detail in chapter 14, means the concept of attention should be treated with caution.

Leslie's (1993) account, on the other hand, is modular. He suggests that common-sense psychology is mostly implemented by a 'theory of mind mechanism', itself divided into two components, one concerned with goal directedness, and the other with mental metarepresentation. These correspond, roughly, to Wellman's simple desire psychology and the later augmentation of beliefs. These components mature at different stages as the adult's common-sense psychology is formed.

Whichever model turns out to be most correct, it must account for the dramatic changes in a child's common-sense psychology between the ages of two and a half to four, shown in figure 3.2. Generally speaking, though, all the psychological models agree that an ability to ascribe beliefs to others is not completely innate. And, as some suggest, there are differences depending on whether the belief is being ascribed to an agent or an object.

The animate and the inanimate: agents and objects

Overall, the difference between Wellman's and Baron-Cohen's theory on the one hand and Perner's on the other might seem rather superficial, but for my purposes it is fundamental. Although both take the theory formation view that children develop a first person and a third person understanding side by side, there is a deep disagreement about the nature of common-sense psychology.

For Perner (1991; 1993) common-sense psychology is really a manifestation of a general representational faculty that is domain independent; the differences between agents and objects are a result of the kinds of representations that are needed by any theory formed to explain them, not due to any fundamental categorical distinction between them. According to this model there is no true distinction between common-sense psychology and common-sense physics. On the other hand, for Wellman (1991), Leslie and Roth (1993) and Baron-Cohen (1993), there is a fundamental distinction between the psychological and the physical. The representations involved in theories that are formed for common-sense explanation are not domain independent; different kinds of representations are used for agents and for objects.

This is an important difference when it comes to building models. On the whole, I think that there is much to be said for Wellman's and Baron-Cohen's stance on this issue, although there is much to be said for Perner's model of the stages of development in a common-sense psychology. I take this view because I think that while there is an important sensitivity to the behaviour of others in Perner's theory, the evidence is that there is some fundamental, almost categorical, distinction between agents and objects. Only this seems able to explain a number of psychological phenomena, for instance naive probability (Humphrey, 1976) and anthropomorphism (Caporael, 1986). And in some ways, even Perner's theory seems to hang on such a distinction; the innate disposition to attend to the behavioural expression of other people's mental states also seems to bring out something of the same difference between agents and objects. I will return to these issues in the second part of this thesis.

Certainly, there is evidence that even very young children interpret agents and objects differently. For example, Meltzoff (1995) found that there is a significant difference between children's ability to imitate actions carried out by a person and the same actions carried out by a machine—well before these children can pass a false belief test. Ricard and Allard (1993) also found that children less than a year old reacted differently to people and to inanimate objects, with unfamiliar animals coming somewhere between. Finally, Gergely, Nádasdy, Csibra, and Bíró (1995) found that 12-month-old children can take the intentional stance to an agent, but only for the goal-directed behaviour that corresponds roughly to Wellman's (1990) simple desire psychology. So for the purposes of this thesis, I will take it as entirely plausible that there is a psychological distinction—although not necessary a clear or consistent one—between agents and objects, and that this is reflected in the distinction between common-sense psychology and common-sense physics.

A brief digression: metaphors and models in psychology and science

There is another part of psychology that also has an important connection with common-sense psychology, although this connection is less obvious than the developmental research into theory of mind. As I've discussed in reviewing the psychology of common-sense psychology, there are two main competing theories: the theory theory which suggests that we learn how to ascribe

mental states to ourselves from observing others, and the simulation theory which suggests that we learn to ascribe mental states to others from observing ourselves. In both cases, there is a special kind of analogical inference going on. In the theory case it is something like 'I believe that I have a mind, because I am like you and you have a mind', where in the simulation case it is more like 'I believe that you have a mind, because you are like me and I have a mind'. Both depend on a similarity between the individuals ascribing and being ascribed mental states. This seems to indicate a special connection between metaphor and common-sense psychology.

There is some psychological evidence for this, but this possible connection has not been extensively studied to date. Nevertheless, the notion of similarity is at the heart of much of the work on metaphor in cognitive science (e.g. Ortony, 1979; Tversky, 1977), and this parallels the importance of similarity in common-sense psychology (Eddy *et al.*, 1993); I will discuss the significance of similarity in much more detail in chapters 7 and 8.

A second, and perhaps clearer, connection is through the study of pretence. Leslie argues that pretence is at the heart of common-sense psychology. Leslie (1987) points out that there is a close correspondence between the three key properties of propositional attitudes and the different kinds of pretence; that is, an imaginary object corresponds to non-entailment of existence, object substitution corresponds to referential opacity, and a pretend property of an object corresponds non-entailment of truth.[1] For example, when I pretend that a banana is a telephone, that does not mean that I universally think of bananas as being telephones, only within the context of my pretence, and this context dependence is also typical of referential opacity in modal logic representations of propositional attitudes. Leslie further argues that this correspondence is not a coincidence, but happens because, underneath, the mechanisms of pretence are those of the "theory of mind mechanism" which implements people's common-sense psychology. The banana is clearly metaphorical; by pretending that a banana can stand for a telephone, the banana is being used as a metaphor for a telephone. And perhaps most significantly, Inagaki and Hatano (1991) argue that

[1] I will discuss the logical properties of propositional attitudes in more detail in the next chapter, when I come to review the work in artificial intelligence on logics of knowledge and belief.

common-sense psychology does give children the ability to use the "person analogy" to predict the behaviour of unfamiliar animate objects; suggesting, following Carey (1985), that common-sense biology develops by extending common-sense psychology to animate objects.

There is another important connection between metaphor and the study of common-sense psychology; this time, it is the role of metaphor in science that is important. Metaphor plays a central role in science, and psychology is no exception. Scientific metaphors are different from literary metaphors in that they are what Boyd (1979) calls "theory constitutive"; that is, they form a real part of a scientific theory. Take, for example, the information processing metaphor, which, broadly speaking, takes people to be like computers and thought to be like information processing. This metaphor is a basic part of a theory because there is no other, more literal, way of making the same theoretical claims. Its role in that theory is that it has "inductive open-endedness" (Boyd, 1979); that is, it sets up a research context in which future study of people and computers may reveal new important similarities and analogies between them. Paradoxically, it may be this use of metaphor that is one of the closest affinities between a scientific theory and the "bedrock" (Clark, 1987) kind of theory characteristic of common-sense psychology. Both are metaphorical and offer an inductive open-endedness.

There is also an important connection between a metaphor and a model. Pribram (1990) describes how scientific progress tends to begin with a metaphor, which is gradually refined as "the original metaphor is transformed into a precise scientific model, a theoretical framework that can be shared by the larger scientific community" (Pribram, 1990). "Every metaphor is the tip of a submerged model" (Black, 1979). All the psychology of metaphor, then, also plays a role in the psychology of the models, and therefore in science.

These connections have serious implications for psychology. If people have inherent tendencies to certain kinds of metaphorical reasoning because of common-sense psychology, and if science is necessarily dependent on metaphorical reasoning, it means that common-sense psychology may affect the interpretation of metaphors and models in science, and in psychology in particular. I

will argue later in the thesis, in chapter 14, that this does in fact happen, that there are indeed systematic biases in our interpretation of metaphors in psychology, and that we need to study these biases scientifically to ensure that we are safe, methodologically speaking.

Summary: the psychology of common-sense psychology

To conclude, the lessons from the psychology of common-sense psychology are mixed. There are a few definite conclusions, for example, we can pretty much rule out Fodor's (1985) claim that common-sense psychology is basically all present from birth. Something is present from birth, to be sure, in all the many different models, but in none of them is it the whole of common-sense psychology, and besides, purely nativist accounts do not match the empirical results; see, for example, Wellman's data in figure 3.2. Wellman's results do show that the development of common-sense psychology follows a consistent, but remarkably complex, pattern.

It is also clear from this review of the psychology of common-sense psychology that a lot hangs on the difference between a first person and a third person perspective, and which of these has primacy. All the first person perspectives reduce, in some sense at least, to a form of egocentric dualism, and therefore to the philosophical 'other minds' problems. The most significant alternative to the composite of first person approaches is the theory formation version of the theory theory (Astington & Gopnik, 1991; Perner, 1991; Wellman, 1991). Instead, it proposes a co-development (or even co-evolution) of the first person and third person perspectives, which escapes the trap of egocentricity. This is attractive for its avoidance of philosophical pot-holes.

Both the two most plausible models to my mind, Baron-Cohen's and Perner's, propose an initial innate sensitivity; in Baron-Cohen's case to other people's attention, and in Perner's case to other people's expression of mental states. There is evidence for both, but I don't think there is necessarily anything contradictory between them. In a sense they are merely different sides of the same coin; sensitivity to attention provides access to the proposition and sensitivity to expression provides access to the attitude, and both are required to develop an understanding of something that resembles propositional attitudes. But both theories seem to stop short; for example, in Perner's (1991) theory of the development of common-sense psychology, the first step on the road to the

child's theory of mind is a sensitivity to the expression of mental states. Here I am suggesting that there is another element at least, another precursor, which is a faculty allowing children to distinguish between things which are agents and things which aren't—that is, between things which can have (and, therefore, can express) mental states and things which can't. This is where anthropomorphism comes into play. As I'll show in the model in chapter 10, this distinction is needed for Perner's situation theory to be operable. Baron-Cohen's (1995) shared attention mechanism may play a similar role, but it is still important to know who to share attention with. Again, something like anthropomorphism is needed.

The immediate contemporary contrast between the simulation theory and the theory theory perhaps isn't so relevant as it seems. The lesson from traditional artificial intelligence is that both have their place, and that both are probably necessary, and Perner's psychological evidence backs this up (Perner, 1994). The problem then resolves to one of combining the two. Even so, the two are very different approaches to predicting reasoning about other people's behaviour, at least in principle, and when it comes to building theories of this competence this is a very significant difference.

As with many other aspects of psychology, there are a lot of conflicting theories and opinions in the study of common-sense psychology. There is simply no consensus on an appropriate theory for common-sense psychology. Samet suggests that this is in large measure because the theories just aren't clear enough, and goes so far as to recommend the use of artificial intelligence simulation techniques to build models. "We actually need something that will work, something that will take inputs and give us reliable outputs—and we don't have such a thing" (Samet, 1993). This is a basic aim of this thesis, to provide just such a simulation, and to use it to explore the extent to which such a simulation can contribute to philosophy, psychology, and artificial intelligence. On this note, then, I will now turn to the last of the three main disciplines with an interest in this area, artificial intelligence, and look at its approaches to an understanding of common-sense psychology.

Chapter 4

Common-sense psychology in artificial intelligence

Introduction

Even in the very early days of artificial intelligence it was clear that for a system to be able to represent and reason about the real world it needed something like human common sense (e.g. McCarthy, 1959). In those early days, though, it meant just that—something that was only *like* human common sense. McCarthy, for example, had originally defined a program to have common sense "if it automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows" (McCarthy, 1959).

McCarthy and Hayes introduced a methodology for studying common sense (Hayes, 1985b; McCarthy, 1959; McCarthy & Hayes, 1969) on the basis of this automatic deduction that looked principally at "epistemologically adequate" representations of common-sense knowledge, separating this from "heuristically adequate" descriptions of the processes which would actually use that knowledge. Generally, this methodology used first order logic or something similar for representing this knowledge, and assumed the existence of an automatic deduction system which would provide the heuristic adequacy given the epistemological adequacy of the representation. Ever more sophisticated representations were tried over and above first order and modal logic (e.g. Minsky, 1981), but the problems of common sense have not yielded (Dreyfus & Dreyfus, 1988; McDermott, 1987). The policy of forgetting about control hasn't proved successful in practice; the assumption that representations can be separated from use of those representations is suspect, and some measure of heuristic adequacy is probably needed from the start.

Artificial intelligence generally breaks down common-sense reasoning into the two broad categories of physics (including spatial and temporal reasoning) and psychology (including knowledge, belief, and planning) although the two are not completely independent. Planning intended actions, for example, requires an ability to reason about time and space. For the purposes of this thesis, though, I will again focus on the psychological aspects of common-sense reasoning.

Part of McCarthy and Hayes' strategy to make common-sense reasoning tractable was to use 'micro-worlds' (Davis, 1992; Dreyfus & Dreyfus, 1988), working with small isolated domains such as the behaviour of different forms of liquid (Hayes, 1985a). Unfortunately, this merely postpones many of the problems: all the untidy bits of a domain can be ignored until we get to the bigger problem of putting together all the different domains. This methodologically simplifying assumption hasn't been proven; nobody has succeeding in putting together any substantial domains—in fact, hardly anybody has even tried.

A notable exception to this is Lenat *et al.*'s (1986) CYC project, which is probably the first and certainly the most ambitious attempt to date to break out of the use of micro-worlds in artificial intelligence. CYC's goal is to create a web of connections between common-sense concepts that can act as a large-scale background common sense, and even as a tool for bridging the gaps between domains. But CYC, too, is explicitly criticised by McDermott (1987) for making the same methodological assumption that representations and the use of those representations can be separated.

So there has been an accumulation of methodological criticisms of artificial intelligence work in common-sense reasoning, both criticisms of the kinds of representation used in artificial intelligence, and even whether the whole concept of representation is appropriate. All this has led some to claim that common sense has been the biggest failure of artificial intelligence (e.g. Dreyfus & Dreyfus, 1988). There is an element of truth in this, but I would argue that even so, artificial intelligence has contributed substantially to a better understanding of the scale of the problem. In this chapter I will review some of the ways that artificial intelligence has shed new light on the problem of common-sense psychology

Common-sense psychology in artificial intelligence

Within artificial intelligence, research on common-sense psychology has mostly concentrated on explicit representations of attitude psychologies, and is firmly within the belief-desire-intention model of mental states (e.g. Cohen & Levesque, 1990; Rao & Georgeff, 1995). In fact, most of the work has concentrated on knowledge and belief on the one hand, where knowledge is interpreted as justified (that is, true) belief, and on planning on the other, with its representation of desires and intentions. The axiom that knowledge must always be true can be represented:

$$knows(agent, \phi) \Rightarrow \phi$$

These representations of attitude psychologies are almost invariably defined with purely syntactic operations, such as the inference rule *modus ponens*, using a variety of modal logic. I will give a more detailed analysis of the representational approaches to common-sense psychology in the next section, but here it is appropriate to point out that this restriction to syntactic manipulation is consistent with traditional representational artificial intelligence—and is therefore subject to all the problems of intentionality in the relationship between syntax and semantics discussed in chapter 2.

At the core of the work on common-sense psychology in artificial intelligence is the representation of knowledge and belief. While an ability to represent and reason about another agent's beliefs is undoubtedly a key part of a common-sense psychology, it is not the whole story. Hobbs for instance, argues that "mutual belief is the foundation on which a common-sense theory of the social world must be built" (Hobbs & Moore, 1985). Mutual belief is far more complex than simple belief (Genesereth & Nilsson, 1987; Hobbs & Moore, 1985), although whether it can be formalised at all in the ways that these studies suggest is, I think, somewhat dubious.

There is a notable and important exception to the work on attitude psychologies, and that is with situated systems. A situated system relies on its environment to provide a semantic grounding, and largely avoids the use of representation altogether (Brooks, 1991a). Even so, some situated

systems can still be characterised as using attitude psychologies; indeed, Rosenschein's (1985) situated automata approach is derived from a possible worlds interpretation of propositional attitudes.

But common-sense psychology is still in its very early days; perhaps the theories that have been developed to date are just not rich enough yet. "We attribute to people not only beliefs but values and emotions as well. We describe them in terms of tendencies and character traits" (Hobbs & Moore, 1985). It is far from clear that the belief-desire-intention nexus is sufficient to provide anything like a complete description of people's common-sense psychology (Dennett, 1987; Samet, 1993). But before analysing whether it is logically possible to run as far as a complete common-sense psychology, perhaps we should try a short walk through the approaches that do exist.

Artificial intelligence approaches to knowledge and belief

There are three principal approaches to representing and reasoning about knowledge and belief. All differ quite fundamentally from the requirements of common-sense physics, in that common-sense psychology requires referential opacity—two symbols denoting the same object cannot simply be substituted for each other. In practice all three approaches use modal logics or something similar to achieve the effect of referential opacity.

The first approach to representing knowledge and belief is to use an augmented modal logic (e.g. Cohen & Levesque, 1990) with additional operators on sentences. These additional operators provide referential opacity by blocking substitution. That is

knows(stuart, 9 > 7)

does not imply that

knows(stuart, number of planets > 7)

even when the number of planets really is 9. The substitution of *9* for *number of planets* isn't valid except in the right context, that is, when:

knows(stuart, number of planets = 9)

Just because the identity holds in reality doesn't imply that the agent *stuart* knows that the identity holds, so the inside of the *knows* expression and the outside of the *knows* expression form different contexts, and inferences are only possible within whole contexts, not between them. That is, the *knows* modal operator is referentially opaque.

A second approach is to use a first order (nonmodal) logic, but with knowledge and belief represented by predicates taking a quoted sentence argument (Haas, 1986). Again this blocks substitution in the different context; this time because the sentence is a string rather than a logical expression. The inference rules which apply to quoted strings are different from those which apply to forms as a whole. Superficially this is similar to a modal logic, but they are radically different in their underlying semantics. In this interpretation, the same propositional attitude would be represented as:

knows(stuart, "number of planets > 7")

The third approach is to represent knowledge and belief through sets of possible worlds (Moore, 1985). This time substitution is blocked because substitutions belong in different worlds. Possible worlds semantics work well for knowledge, particularly nested knowledge statements, but they are poor at handling beliefs (Cohen & Levesque, 1990; Genesereth & Nilsson, 1987). But a possible worlds approach has one strong advantage: it lends itself to dealing with an environment as well as an agent's mental states (Rosenschein, 1985). A possible worlds approach makes it easier to handle the connections between attitudes and the real world.

All these different representations suffer from many of the same shortcomings. The first of these is associated with observation. A typical observation axiom is:

$$\phi \Longrightarrow knows(agent, \phi)$$

This axiom says that *agent* knows ϕ when ϕ is true of the world. This particular observation axiom is never really true because things which are true of the world might not be observable by a given agent: false beliefs like those of chapter 3 break this axiom, and more complex observation rules

are always required in practice (e.g. Davis, 1988). Also, this axiom is only descriptive of an agent observing the world, it provides no causal account. It is worth noting that this observation axiom is symmetrical with the 'knowledge must always be true' axiom; axioms in these roles describe the 'world to mind' and 'mind to world' links required by an agent.

Intentionality is another problem for all these representations of knowledge and belief. The purely syntactic approach of logic is normally taken to offer only 'as if' intentionality derived from the people who used it, not the true intentionality that is required to properly connect an agent's propositional attitudes to its environment. As I claimed in chapter 2, although there are problems with a categorical distinction between original and 'as if' intentionality, in practice this distinction is often approximately correct. The intentionality will generally be 'better' if the causal connections between symbols and objects in the real world are visible to observers. Again we regress to the importance of the observer.

But perhaps the most telling criticisms are simply those of the incompleteness of these logical approaches. While they all seem to handle knowledge well, even belief is pretty difficult in the possible world semantics, and other attitudes, such as desires and intentions, have proved far more elusive. Mapping attitudes into operators is not a trivial exercise (Cohen & Levesque, 1990). Modal operators are excellent for handling a logical theory, but this presupposes that the theory of common-sense psychology is logical. While this may be approximately true for knowledge, it seems dubious for the more specifically human mental states, some of which (moods, for instance) can't easily be thought of as attitudes to anything at all. It is perhaps for this reason that formalisation of common-sense psychology in this area has been almost entirely restricted to knowledge and belief, the least subjective and most obviously attitude-like aspects of an agent's mental states.

Finally, I will conclude by suggesting that in no way do any of these logics provide any clues to the distinction between the physical and the mental. Many of the same logical properties needed for beliefs are needed for reasoning about time, for example, and the logics of knowledge and belief

are mostly the same as the standard modal logic S5. None of these logical systems, therefore, provide anything other than a simple category distinction between things which are agents and things which aren't—exactly the kind of distinction criticised so strongly by Sloman (1993).

To summarise, the role of logical languages for representing and reasoning about knowledge and belief seems to be exactly similar to that in Hayes' (1985b) descriptive models of naive physics. In so far as it is descriptive, it is a valuable tool; but the clues it offers to the actual mechanisms which underlie common-sense psychology are very limited.

Artificial intelligence approaches to desires and intentions

The other aspect of artificial intelligence approaches to common-sense psychology is concerned with desires and intentions. This usually synthesises the traditional artificial intelligence methods of planning with common-sense reasoning about knowledge and belief. An agent's actions both depend on its beliefs, and can change those beliefs (Moore, 1985). An agent's desires are identified with the goals of a planning system working with those beliefs (Cohen & Levesque, 1990). But there are differences from traditional approaches to planning, however. Because with common-sense reasoning an agent is no longer assumed to be omniscient, desires and intentions must also become referentially opaque (Davis, 1992), and need representations similar to those of knowledge and belief. Methodologically, therefore, artificial intelligence's approaches to common-sense reasoning about desires and intentions are similar to those of reasoning about knowledge and belief.

An important point raised by reasoning about desires and intentions is the role of an agent's actions. Actions are symmetrical with observations—they are another part of the connection between the agent and its environment. It is essential to distinguish between actions and plans (Moore, 1985; Suchman, 1987) just as it is between observations and inferences (Genesereth & Nilsson, 1987)—and for the same reason. The hidden complexities of observation axioms are paralleled by equivalent hidden complexities of action; for example, an action may fail to achieve what the agent intended it to do.

The axioms which describe links between the world and the mental states of an agent are related to the causal links in figure 3.1. This reveals an apparently important gap—there is no equivalent desire, or goal, axiom. This is probably inevitable in this methodology, in that there doesn't seem to be anything obvious in the real world to relate to desires as events do to observations and actions. And, as I've already hinted, representational artificial intelligence has already proved rather less appropriate for handling the desire aspects of the belief-desire-intention nexus.

This difficulty lies at the surface of an important problem. Goals contain an inherent aspect of control, so the problem of methodologically ignoring control issues in traditional representational approaches to common-sense reasoning in artificial intelligence will inevitably tend to make goals more difficult than beliefs and actions. To be sure, goals can be added (Cohen & Levesque, 1990) but they do, in part at least, run counter to the straight use of deductive reasoning.

Distributed artificial intelligence

So far, this analysis has concentrated on the psychological structures involved in a single agent. This is only part of the story: the whole point of common-sense psychology is to predict and reason about another agent's psychological states, so the single case is too simplistic. In recent years there has been a dramatic growth of interest in distributed artificial intelligence, in which many agents act in a coordinated or cooperative manner. Some of the motivations for this are technical, associated with advances in distributed computing. But perhaps more interesting are the social motivations. As computers become more integrated with people's everyday lives there is a move towards a "grand collaboration" (Kay, 1990) of people and agents, where agents need to be able to collaborate effectively both with each other and with people.

In distributed artificial intelligence agents need to interact and negotiate. Much of this work has its roots in Austin's (1962) work on "speech acts"—communicative acts which one agent can use to affect the behaviour and mental state of another. For example, one agent's saying 'the moon is made of green cheese' may be an attempt to get another agent to believe that the moon is made of green cheese—an attempt to create a propositional attitude in another agent—and therefore to indirectly affect its behaviour. Communication becomes a kind of action, a speech act is 'for'

something—for acting on another agent's mental states. Searle (1969) then combined Austin's speech acts with Grice's (1957) analysis of intentions to provide conditions for the performance of speech acts. This unified formalism provided distributed artificial intelligence with a good enough model of communication to be able to combine communicative acts with reasoning about beliefs, desires, and intentions (Cohen & Levesque, 1990). Ultimately, though, distributed artificial intelligence still suffers from the same shortcomings as conventional artificial intelligence in its treatment of common-sense psychology.

But for the purpose of this thesis perhaps the most important role of distributed artificial intelligence is as a modelling technology (e.g. Wesson, Hayes-Roth, Burge, Stasz, & Sunshine, 1981). Just as cognitive psychology gained from artificial intelligence as a technique for modelling individual minds, social theories can gain from distributed artificial intelligence as a technique for modelling the interactions between individual minds. It is in this spirit that the models in the third part of this thesis should be understood. Because common-sense psychology plays a fundamental role in social interaction, distributed artificial intelligence has a lot to offer as a modelling framework.

Representational artificial intelligence revisited

This review of the representational approach to models of common-sense psychology shows that it does work well within reasonable limits. It provides a relatively complete and self-consistent model of the belief-desire-intention nexus, and it integrates beliefs, desires, and intentions with perceptions and actions, for individuals and for societies of interacting agents. But on the whole, while representational artificial intelligence has been relatively successful, its success conceals some deeper issues. Although representational artificial intelligence provides almost everything that Fodor (1985) asks of a representational theory of mind, there are problems that remain over and above the problems of semantics that Fodor believes divide artificial intelligence from psychological validity. In the next two sections, I will examine two of these in more detail.

It is also worth noting that one obvious challenger to representational artificial intelligence—connectionism—hasn't really been a force of significance in common-sense psychology. Indeed, it was from the high ground of common-sense psychology that Fodor and Pylyshyn made their "systematicity argument" attack on naive connectionism (Fodor & Pylyshyn, 1988). A full exposition of their argument is beyond this thesis, but it is based on the compositional and generative aspects of propositions mentioned in chapter 2, and the problems of implementing these aspects with connectionist systems. Briefly, they argue that because propositional attitudes can nest as beliefs about beliefs and so on to any depth, no connectionist system can actually manipulate them without being, underneath, just a connectionist implementation of a conventional symbol processing system and, therefore, the apparent virtues of connectionism may well be illusory.

Despite these problems, in short, representational artificial intelligence is the only approach which has to date measured up as capable of providing a wide-screen yet broadly valid model of common-sense psychology. This does not mean that it should be accepted as the only possible approach—even its adherents agree that important lessons can and should be learned from other approaches (Hayes, Ford, & Agnew, 1994). And yet this very completeness, one of its most powerful and attractive features, is its Achilles' heel.

The frame problem

One of the problems with reasoning about the real world is that it is impossible to put a 'frame' around the knowledge that is required to deal with a particular situation. If an agent is to build a real-time plan for acting in a real world, it needs to disentangle the relevant implications of its actions from the irrelevant ones. This appears to be impossible: there is a finite time available to produce a plan, but an infinite number of possible implications to investigate; so time will run out before the agent can build a proper and correct plan. In a micro-world, of course, this is not a problem: the micro-world can be defined so that exceptional situations can be foreseen—the micro-world itself provides the frame. The problem is only one of the real world, where there doesn't seem to be any such frame, and the gaps between the different frame assumptions will only be revealed when attempting to put together all the different micro-worlds.

This is the problem that McCarthy and Hayes (1969) originally called the "frame problem". It is, in effect, the problem of finding all the relevant information to the problem in hand, and not getting bogged down explicitly ignoring all the irrelevant information. Dennett (1984) sees this as one aspect of a fundamental new philosophical problem—a general epistemological relevance problem—one which Hayes calls the "whole pudding", but following Dennett I will appropriate the term 'frame problem' for this general relevance problem.

The traditional solution to the frame problem is to get an agent to ignore things which it decides are irrelevant, but this leads to another part of the whole pudding, a closely related epistemological relevance problem, the qualification problem (McCarthy, 1977). McCarthy illustrates the qualification by means of the 'cannibals and missionaries' problem. Normally, of course, the solution to the cannibals and missionaries problem doesn't involve a bridge half a mile upstream—although this isn't explicitly excluded by the description of the problem. The problem hasn't been qualified by ruling out all the unmeant alternative possible solutions. There is no limit to the possible solutions 'outside' the problem, but people don't usually resort to them (for an example of someone who did, see "The Barometer Story", Calendra, 1964).

On the other hand, it is clear that people do solve the frame problem in practice—or at least that they appear to do so. How they do it, we don't know; but do it they must. It is also clear that the solution must be heuristic rather than exact, in that people cannot actually take into account everything that might be relevant, only those things that are relevant enough to be considered. But the precise nature of the heuristics is still very elusive.

There is a clear moral from the frame problem. If it is not possible to represent the real world complete as it is, then perhaps the next best thing is to use the real world itself instead of a representation to fill in the gaps. This is the approach of situated systems. In extreme cases, all representations in a situated system are entirely notional; that is, there are no explicit representations, only 'as if' ones which an ascriber might ascribe to the system. In milder cases, situated systems can be combined with explicit representation. But in all cases of the situated approach meanings are not 'in the head' (or even 'in the computer') but are part of the interaction between an embodied system and its environment.

Situated approaches to common-sense reasoning

Situated artificial intelligence is a radically different approach to common-sense reasoning. Justifications for situated systems are rarely framed in terms of common sense, but there is a strong connection between the two, in that both are intended to deal with a complex real world. In a situated system, instead of representing the world, the world is used instead of or as part of the representation (Brooks, 1991b). Because of this divergence from explicit representation, if there is anything like an attitude psychology, it will have to be notional rather than propositional or sentential. Harnad (1990) even argues that an element of situatedness in some form of sensorimotor grounding, for example, is required to provide intentionality, as I've suggested in chapter 2. Situated systems have generally proven more far more successful at handling complex physical environments, weight for weight, than their representational cousins (Brooks, 1991a; 1991b), although if taken to the limit they do tend to throw the baby out with the bath water (Hayes *et al.*, 1994; Sandberg & Wielinga, 1991).

A deeper criticism of this approach is that extending situated representation to the psychological domain hits a major block in that agents are opaque; a person's mental states are not 'there' for grounding in the same way that physical objects are. This is not intended to be an argument for a causal account based on propositional attitudes—I've discussed my misgivings on that in chapter 2—but it is an argument for something that can behave like propositional attitudes. In particular, something like referential opacity is required. A fully situated approach to representation, such as that of Brooks (1991b) or Rosenschein (1985) does hit problems in common-sense psychology, in that multiple agents in the same environment add complexity. Consider the simple situated system in figure 4.1. In this system, there is a single thermostat in an environment, and we can take B, which is a situated representation of the temperature of the environment, as a 'belief-like state' and D, a situated action on the environment, as a 'desire-like state' (Sloman, 1993). Now consider a multiple thermostat version of the same scenario, shown in figure 4.2.

In figure 4.1, it is clear that there is a relationship between D, B, and the temperature in the environment. But in figure 4.2, the relationship is far more complex. The individual B_is and D_is no longer refer only to the environment, but also indirectly to the B_js and D_js of the other thermo-

stats, and the correlation between the B_js varies according to the difference in the D_js. In this example, for thermostat A the internal states of thermostats B and C constitute part of its environment, yet they cannot be directly perceived by thermostat A. This means that although it is possible to talk about A's environment including B and C, this glosses over important aspects of the system as a whole.

This shows the difference between common-sense psychology and common-sense physics, and it highlights the problem of grounding in psychological and social systems. When there is no possibility of the kind of direct perception that Harnad (1990) and Brooks (1991b) argue is necessary for grounding, we are reduced to exactly that kind of indirect grounding that would be called 'as if' intentionality. We know, as observers of the system, that thermostat A's state depends on the state of thermostats B and C as well as that of the environment, but thermostat A cannot be

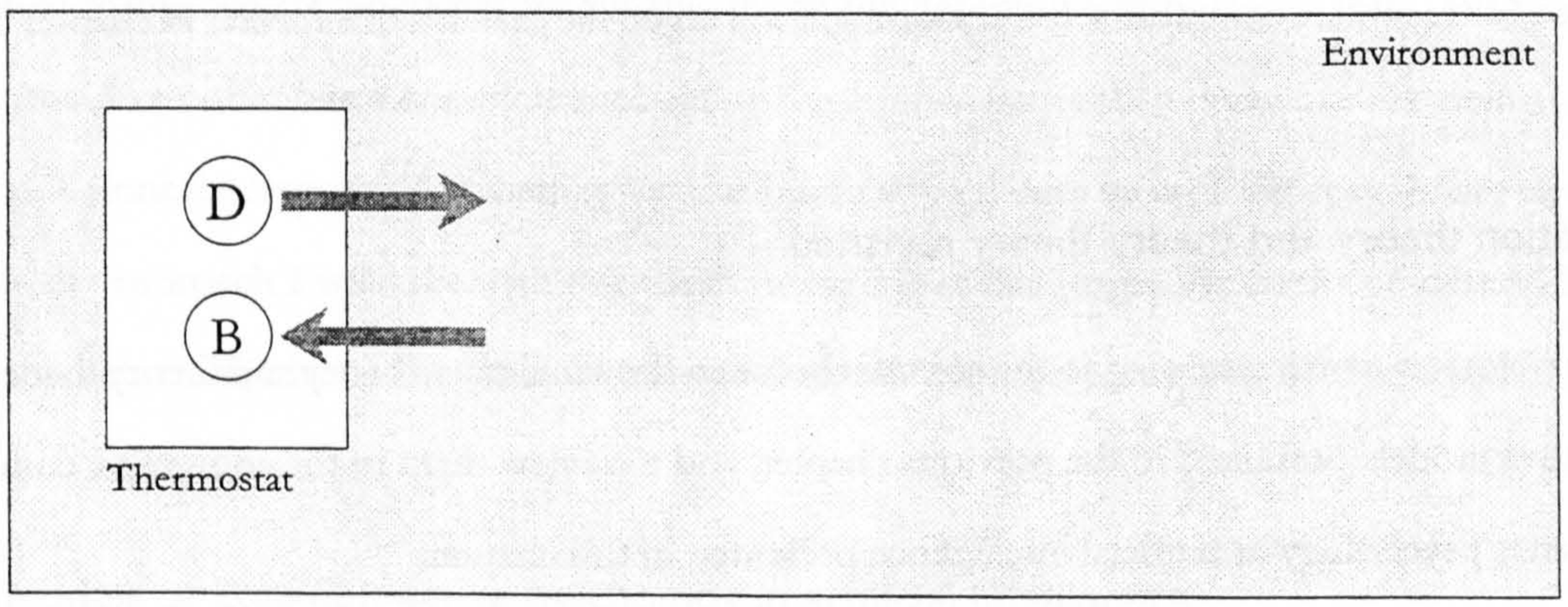

Figure 4.1. A simple situated system

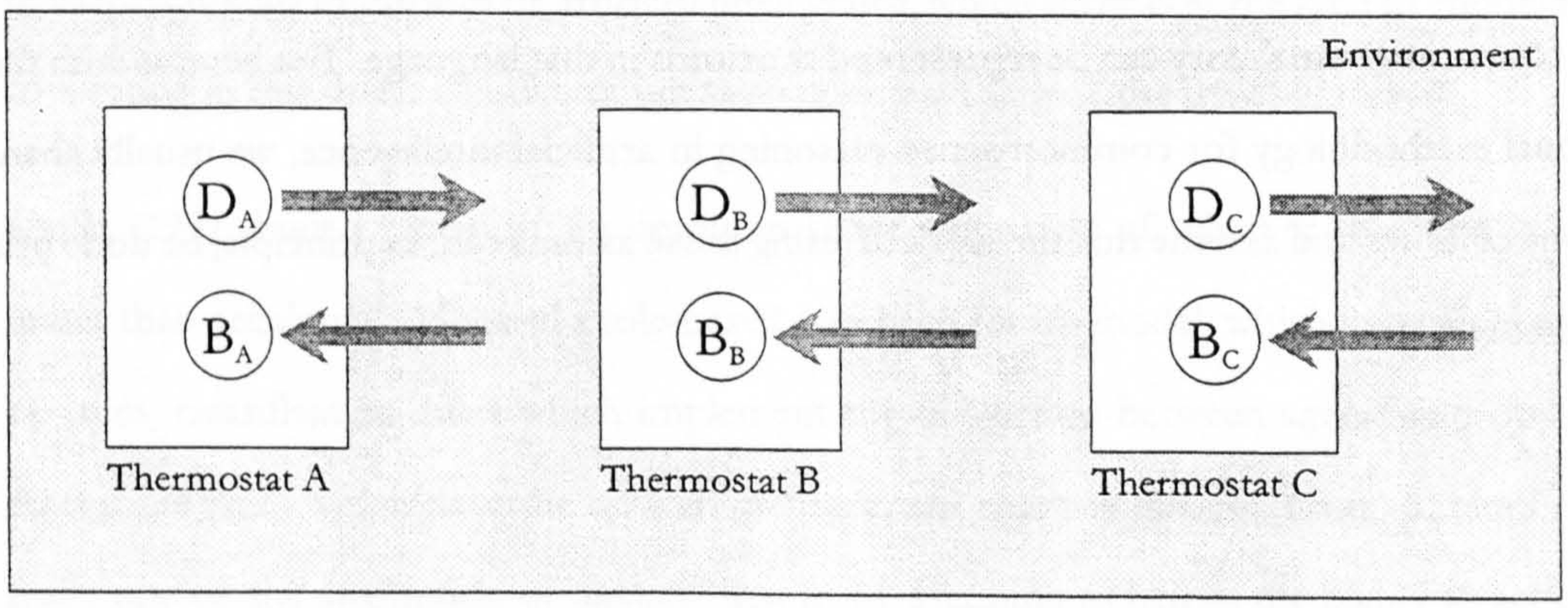

Figure 4.2. A less simple situated system

grounded in these states because it has no direct access to them. Since this is exactly the kind of grounding that situated systems are intended to avoid, for psychological and social systems the situated approach can only ever be partially successful.

Another situated approach to common-sense reasoning, that of Steels (1988), explicitly suggests replacing propositional representations, at least implementationally, with analogical mapping. As with Brooks' work, this seems appropriate for physical environments, and, again as with Brooks' work, it shows no sign of being extensible to common-sense psychology.

So while situated systems are able to deal with complex physical environments—often far more complex environments, weight for weight, than similar traditional representational systems, there is no sign that they can be generalised to psychological or social environments. And paradoxically, it is the virtue of the referential opacity of modal logic that it can handle the kind of indirect representation needed for these psychological and social worlds. Properties of both seem to be required for complete common-sense reasoning; I will argue the case for this further in chapter 6.

Simulation theory and theory theory revisited

At this point it is worth returning to the contrast between the simulation theory and theory theory alternative models presented in the previous chapter, and to review them in the context of common-sense psychology in artificial intelligence presented in this chapter.

Representational artificial intelligence is clearly closely related to the theory view. In this way, a common-sense theory of liquids (Hayes, 1985a) can be represented in a logical language. The laws of the common-sense theory can be represented as axioms in that language. But here, as with the traditional methodology for common-sense reasoning in artificial intelligence, we usually abandon control issues and assume that the work of using those axioms can, in principle, be done by a fancy theorem prover.

The simulation theory, on the other hand, imposes different requirements. Simulation has something in common with look ahead in game playing. And like efficient game playing, the success of the strategy depends on the quality of selection between different alternative possibilities. This selection is, of course, a control issue, and, therefore, while the traditional methodology for common-sense reasoning avoids control issues, these issues are still important to the simulation theory.

Deductive reasoning, then, can provide a useful description of behaviours generated either by a simulation or by a theory. On the other hand, if we actually need to run a simulation, deductive reasoning with its complete absence of control description just isn't powerful enough. In fact, if we accept that heuristic aspects are an important part of a theory for generating behaviour, then deductive reasoning isn't powerful enough for that either.

So while deductive reasoning appears not to be enough for proper implementation of either a simulation theory or a theory theory, it is enough for a description of both. In this spirit, we are free to use logical languages within this thesis for descriptive purposes at least, but we should provide a rather more complete account of the control issues than deductive reasoning can provide alone, irrespective of whether we take a simulation theory view of a theory theory view. This is the approach I will take with this thesis: using logical languages for clarity of description, but supplementing them when they omit control or other important aspects of description.

Models of common-sense psychology in artificial intelligence

Before concluding this review, it is worth reviewing a few more concrete models of common-sense psychology in the area of artificial intelligence, which show how the kind of modelling I am advocating in this thesis can contribute to an understanding of the issues involved.

Shultz (1991) built a model of the common-sense judgement of whether an action is intended rather than accidental. He used a rule-based approach for his model, which comprises two kinds of rules, classification rules which implement the distinction between agents and objects, and diachronic rules which describe how an environment changes through time. Shultz's model is really one of the attribution of agency; that is, of ascribing to beings the ability to act or move

autonomously without external causation. Shultz suggests that agency is a primitive aspect of common-sense reasoning underlying both animacy (whether things are alive or not) and animalness (whether things are animals or not.)

Shultz's model makes a number of important points. First, it shows that a relatively simple artificial intelligence language (in his case, OPS5) can be used to build quite a sophisticated model, and that this model can provide feedback into other branches of psychology. Second, it allows the developmental aspects of common-sense psychology to be explored through different mechanisms of rule creation and rule modification. On the deficit side, Shultz's model is still a micro-world, and a quite a small micro-world at that—being restricted almost completely to the explanation of autonomous movement. Also, Shultz's distinction between agents and objects is circular and based purely on the capacity for autonomous movement, which, although it is undoubtedly an important factor, it is certainly not the only one (Caporael, 1986). Nevertheless, Shultz's project has much in common with the work in this thesis, both methodologically and theoretically; namely, it shows that descriptive model of common-sense psychology is possible, and that it can contribute to scientific understanding in this area.

Another model worth describing in a little detail is Davis' model of common-sense observation. Davis (1988) draws his model from robotics; it was intended to provide a better observation axiom by providing some guidelines on the inference of what an agent can believe from its situation and its attention. So, if there is an obstacle which prevents an agent from perceiving an event—Davis' example is shown in figure 4.3—another agent can use this to infer that the agent is not aware of the event. Although this is probably sufficient for passing the standard false belief

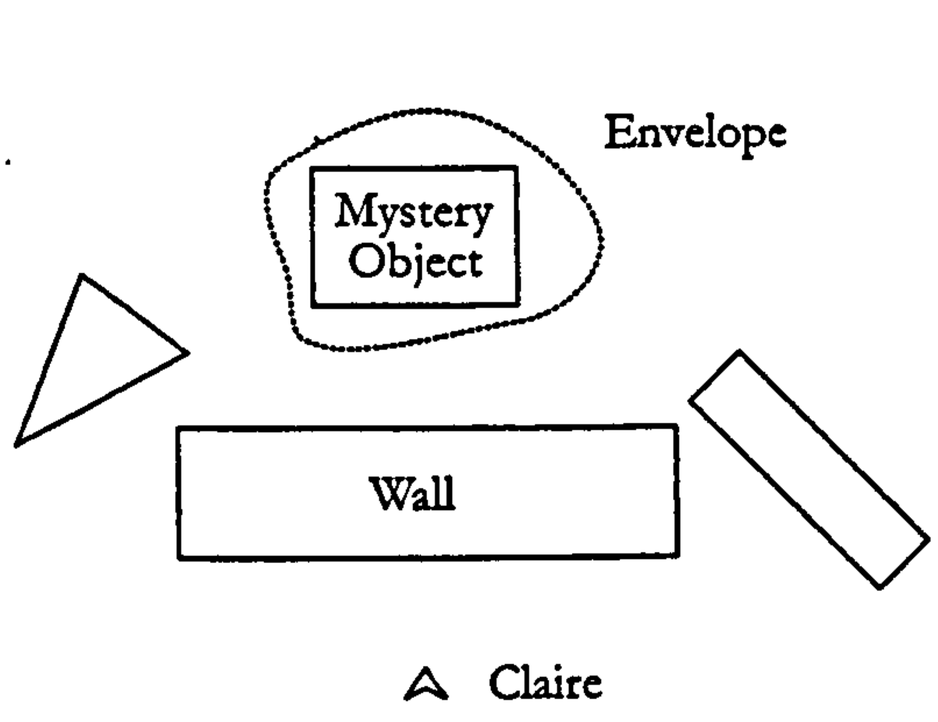

Figure 4.3. Inferring an agent's perceptions from its situation (after Davis, 1988)

test described in chapter 3, it does not seem to have been used as a model for it, and it ignores all the developmental issues raised in the previous chapter. In effect, Davis' model dives in at full adulthood. But while Davis is aware of the full import of his approach, his model is restricted to a fairly simple micro-world, and it again provides no clues to the distinction between agents and objects. Nevertheless, for the problem of inferring ignorance, or false belief, Davis' model is rather elegant.

Summary: common-sense reasoning in artificial intelligence

There are two dominant strategies for common-sense reasoning in artificial intelligence: representational and situated. The representational approach has a longer history and apparently greater flexibility. The situated approach arose as a response to problems in the representational approach, some philosophical (e.g. the 'frame problem') and some practical—building robots with even very simple behaviour seemed just too hard.

But there are a number of arguments which show that neither approach alone is sufficient to meet the requirements of a fully-fledged common sense—a purely representational approach is victim to the frame problem and a purely situated system cannot be grounded in states that it cannot perceive directly.

A second problem is that common-sense reasoning in its representational mode has been dominated by a carelessness regarding psychological validity. The pioneer of the field, McCarthy, definitely views artificial intelligence "as a branch of computer science rather than as a branch of psychology" (McCarthy, 1988). Psychological validity was never a goal of the logician's approach. The conflict between logical and psychological approaches to artificial intelligence is not a new one (Israel, 1985; Kolata, 1982), but there is little to add to this controversy beyond what has already been written (e.g. Israel, 1985; McDermott, 1987; Minsky, 1985). Nevertheless, since this research is intended to be in cognitive science rather than 'hard' artificial intelligence, psychological validity is a principal goal for the purposes of this project—and therefore the psychological issues reviewed in chapter 3 must be applied where possible.

This should not necessarily be taken as an argument against the use of propositional attitudes or even deductive reasoning. Although it seems exceedingly improbable that there is anything corresponding to sentences in logic 'in the head', that still doesn't invalidate logic as a descriptive tool. It does, however, suggest that control issues and heuristics are important, despite McCarthy and Hayes' methodological assumptions—especially for the simulation theory—and, therefore, we do need to be more explicit about these control issues.

These lessons are those which are to be learnt from a direct study of common-sense psychology in artificial intelligence. In the next chapter, I will move on to a more indirect study which has some additional and valuable insights to provide; a detailed analysis of a classic example of the ascription of mental states in artificial intelligence, the Turing test.

Chapter 5

Common-sense psychology in the Turing test

Introduction to the Turing test

Turing (1950) considered the question 'can machines think?' but almost immediately threw it away as "too meaningless to deserve discussion" and proposed replacing it with a more empirical test; the test that has since become known as the 'Turing test'. Turing originally derived his test from a party game called the 'imitation game', shown in figure 5.1, which has a human observer trying to guess the sex of two players, one of which is a man and the other a woman, but while screened from being able to tell which is which by voice or appearance. One of the players will probably try to help the observer by being truthful, where the other may try to deceive the observer by pretending to be of the other sex. Turing's proposal was to let a machine take the place of one of the humans and essentially play the same game, shown in figure 5.2. If the observer can't tell which is the machine and which the human, this can be taken as strong evidence that the machine can, in fact, think. Despite its ancestry, the Turing test is not a game, but a serious suggestion that an indistinguishability criterion does empirically mean something about the system under test (Harnad, 1992).

In Turing's proposal, the screening is provided by a teletype link between the participants. Although arbitrary restrictions in modality can be contested (Gunderson, 1971; Harnad, 1991) this purely linguistic modality can be defended, as Dennett (1985) does, by appealing to the scope of language. Even in Turing's version, no bounds are placed on the permitted interaction, so it is possible to evaluate the system's understanding of the real physical world within limits, by asking questions about it. The absence of bounds means that common-sense physical reasoning can easily be investigated in the framework of the Turing test. Especially important, however, is the interactive format of the test; as a restricted version of human social interaction it is likely to prove a good format for the investigation of common-sense psychology.

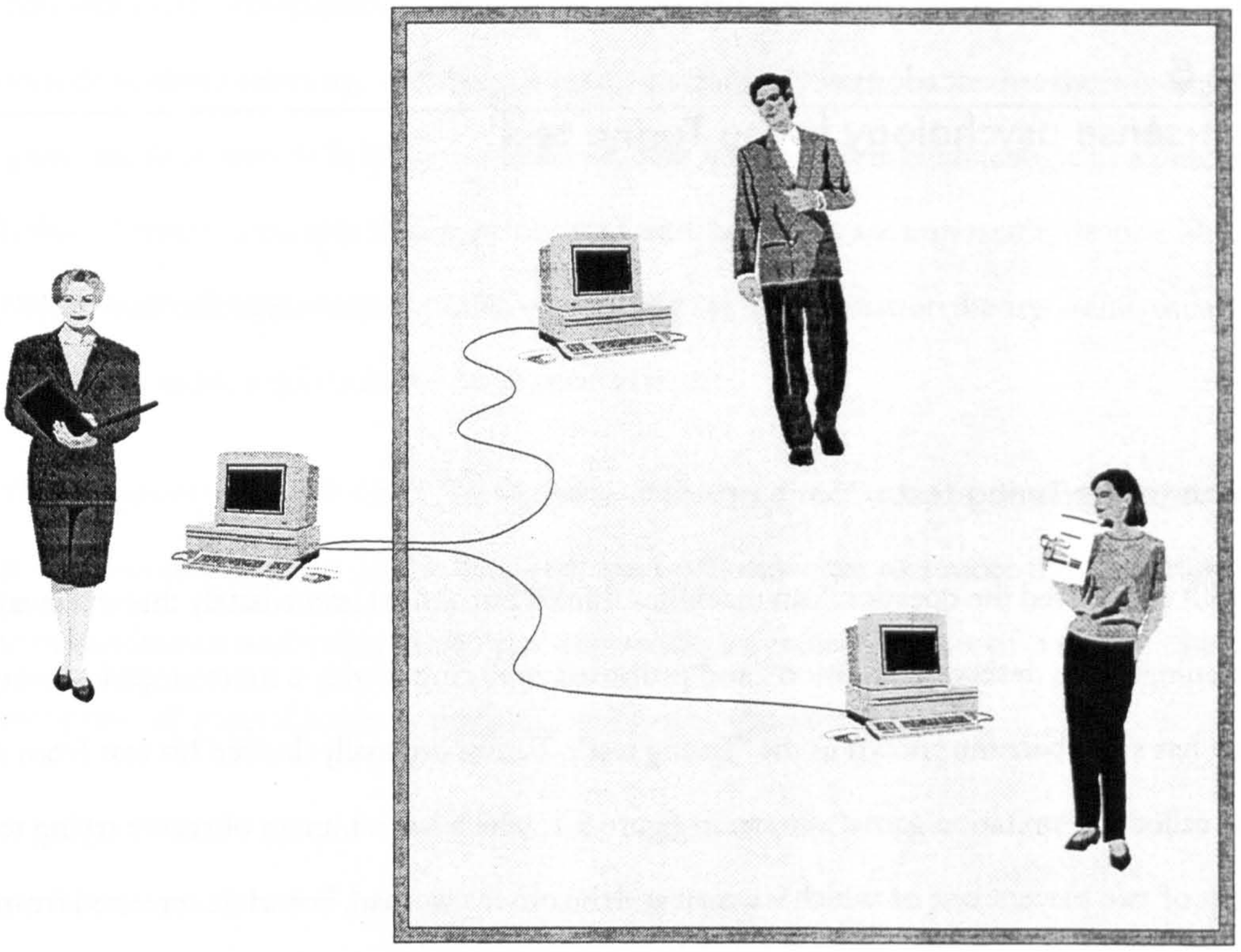

Figure 5.1. The imitation game

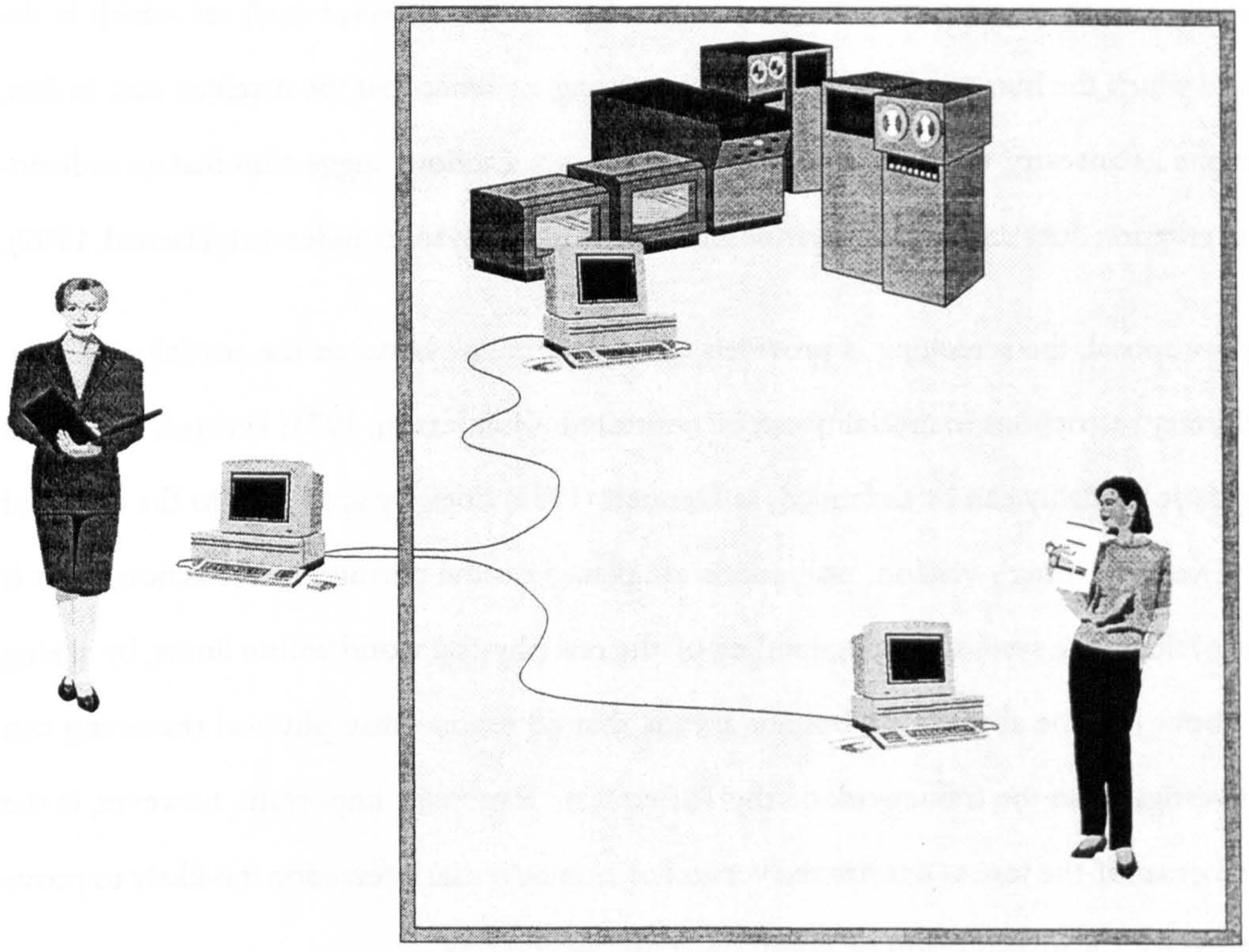

Figure 5.2. The Turing test

But it is not my intention to propose building systems which pass the Turing test. More, it is that the Turing test highlights real world problems when taken as a thought experiment (Dennett, 1985). This chapter reviews the literature on the Turing test in this context, emphasising the important issues that are raised with respect to common-sense psychology.

Problems with the Turing test

Today, few accept the Turing test completely in its original form for one reason or another (e.g. Block, 1981; Gunderson, 1971; Harnad, 1991; Hayes & Ford, 1995; Michie, 1993; Searle, 1980; Whitby, 1996). The most frequent criticisms are that the test is too easy, or that it is behaviourist or operationalist, that it is too narrow, or that it is too shallow (Moor, 1992). For all the criticisms, though, the Turing test has not gone away. As Dennett (1985) says, "there are real world problems that are revealed by considering the strengths and weaknesses of the Turing test". Criticisms of the test, so long as they are not based on too literal an interpretation of Turing's original proposal, can tell us a story about intelligent behaviour in general, not just about how we recognise it.

With respect to the test being too easy, Block (1981) and Weizenbaum (1976) both comment that the test can be passed simply by fooling the judges sufficiently well. In practice, however, seriously critical judges can't be fooled so easily: the test is only too easy if tricks are used and restrictions added, which certainly seems to be against the spirit of Turing's proposal, if not the letter (Dennett, 1985; Harnad, 1992). Similar responses can be made to criticisms of the narrowness (e.g. Gunderson, 1971) and the shallowness (e.g. Searle, 1980) of the test.

Conversely, French (1990) claims that the Turing test is actually too hard because it doesn't just test for intelligence, but for human socially and culturally adapted intelligence, since there is no other kind of intelligence to act as a metric. The Turing test doesn't test for intelligence "in general". Schank and Seifert agree, claiming that intelligence is not a "unitary phenomenon, but a spectrum—a matter of degree" (Schank & Seifert, 1985); understanding, for example, comes in different levels, varying from simply making sense, through everyday understanding, to complete

empathy. They argue that the Turing test is too hard because, in effect, it can only evaluate the highest possible levels of understanding, and has nothing to say about the lower ones. It is, they say, important to not to be too "human-chauvinistic" (Schank & Seifert, 1985).

But here be dragons: if non-human intelligence exists but cannot be detected, then in principle anything could be intelligent, even lettuce and tables. If systems to be evaluated for intelligence are allowed to choose their own interaction modality, then in principle rock boxes (Gunderson, 1971) and thermostats (McCarthy, 1979) might be validly classed as intelligent. The search for non-human intelligence might be a red herring if intelligence only exists to the extent that we humans can recognise it, with the proviso that the behaviour we recognise as intelligent isn't fixed. The role of our ability to detect intelligence is of crucial importance to the Turing test, and the theoretical possibility of radically non-human intelligence is an important one, which I will address in more detail in the next chapter.

Turing didn't make it clear that the test wasn't intended to be an operational definition, so some have interpreted it operationally (e.g. Millar, 1973; Searle, 1980; Michie, 1993) while others have firmly rejected operational interpretations (e.g. Dennett, 1985; Moor, 1976). Searle (1980) specifically describes it as "unashamedly behaviouristic and operationalistic", but on the other hand, Moor (1976) sees its value "not in treating it as the basis for an operational definition but in considering it as a potential source of good inductive evidence for the hypothesis that machines think". Dennett (1985) also opposes an operationalist interpretation of the test. Instead, he implicitly suggests an analogy with legal practice: "Turing proposed that any computer that can regularly or often fool a discerning judge in this game would be intelligent—would be a computer that thinks—*beyond any reasonable doubt*" (Dennett, 1985, original emphasis). But if we want to address the real world issues that the Turing test raises, operationalism is singularly unhelpful.

This legal analogy needs some clarification. Dennett (1985) clearly intends the default answer to be failure: 'not-thinking until proven thinking', which makes nonsense of an operational interpretation when humans fail the test (and they do!) Unfortunately, human psychology isn't on Dennett's side here: there is strong evidence that the common-sense default is, under some circumstances at least, completely the reverse: 'thinking until proven not-thinking' (Caporael, 1986; Weizenbaum,

1966). To be fair to Dennett, though, he too recognises the real tendency to overestimate the intelligence of a system, and recommends that we try to avoid it (Dennett, 1985), which is why he encourages his sceptical default. To this extent at least, the psychological baggage of the observer is critical to the outcome of the test (Collins, 1990).

In response to Gunderson's (1971) criticism of the narrowness of the test, again, a seriously critical judge isn't so easy to fool. While restricting the modality, as Gunderson does in his 'rock box' parody of the Turing test, does have an effect on the judgement, in practice the unrestricted use of language places only very wide bounds on the judgement. Searle's (1980) criticism of the shallowness of the test is more serious and has caused all sorts of enlightening misunderstandings, so I will turn to this in more detail in the next section.

Finally, more recent criticisms of the Turing test (e.g. Hayes & Ford, 1995; Whitby, 1996) argue that it just doesn't help us understand intelligence. The imitative nature of the test is both misleading and scientifically unhelpful. Perhaps even worse, it is actually damaging the field, as the goal of artificial intelligence is often seen as simply passing the Turing test. On the other hand, Hayes and Ford do accept a more humanistic interpretation of the test: through it we are forced to think about what it is to be human. "If this was what Turing meant, then we need not reject it as our ultimate goal" (Hayes & Ford, 1995). I believe it is this human side to the test that may be its biggest asset to artificial intelligence, and it is in this spirit that I will look at it in more detail.

Searle's 'Chinese Room'

Searle's article on the Chinese Room argument (Searle, 1980) only attacks Turing's proposed test in an aside as "unashamedly behaviouristic and operationalistic", but throughout his argument there is an implicit criticism of the Turing test itself. Searle targets his attack explicitly at Schank's work on language understanding (e.g. Schank, 1977) but also notes that the same applies to Winograd's SHRDLU, Weizenbaum's (1976) ELIZA "and any Turing machine simulation of human mental phenomena". Searle makes a number of illuminating points in his argument, but it is perhaps best interpreted as a common-sense point on the futility of trying to pass the Turing test simply by fooling the observer.

Searle's argument is intuitive: he imagines himself in a room, interacting with the outside world through paper messages, but he translates the whole interaction with the system to Chinese, so that he doesn't understand a word of it. Then he imagines himself inside the room simulating by hand a program which shuffles and matches Chinese symbols to generate responses. The room itself corresponds to a system which can pass the Turing test, with the questions and answers in Chinese; while Searle himself, inside the room, doesn't understand a word of Chinese and is just matching patterns and following rules. Since Searle-in-the-room is just obeying rules, even though the system appears to those outside the room to understand Chinese fully, those rules cannot constitute truly understanding Chinese; even though Searle-in-the-room is following the rules, there is no understanding anywhere in the room. He argues, then, that understanding cannot come about by just by following rules and matching patterns. This is unfortunate because that is all a Turing machine can do.

Searle (1980) anticipates and partially rejects a number of standard replies to this argument. The favoured rebuttal from the artificial intelligence community is the "systems reply" (e.g. Hofstadter & Dennett, 1981) which says that although Searle-in-the-room doesn't understand Chinese, the 'system' as a whole (including the room, the rules, and the patterns that Searle-in-the-room is following) really does understand it. Unfortunately, by effectively saying 'well, if we *forget* about what's going on inside, it all makes sense' it has little force as a counterexample. Of course, if we do by some miracle of engineering insanity ever build a Chinese Room that really does understand, then we have to accept the systems reply in some sense, because it must be the system as a whole that has the understanding. So, while we may accept the reply, it isn't particularly informative or helpful, and I will move on to the replies which are more directly relevant to the case for common-sense psychology.

A second reply (e.g. Boden, 1988; Harnad, 1991) is the "robot reply", which transplants the room into the head of a robot. Now, instead of questions being asked, Searle-in-the-robot receives input from sensors connected to the world. The philosophical argument underpinning this reply is that this gives the Chinese symbols on the scripts and rules a grounding (and therefore semantics), Searle-in-the-room can correlate the sensory input for 'red' with the Chinese symbol for

'red', so real understanding is possible (Harnad, 1991). I will return to this issue in the next section. Searle dismisses this reply as not having dealt with the question because the robot system is actually isomorphic to the room system. This is an essential point, and one to which I will return later—one of the more subtle aspects of Searle's thought experiment is that different systems which are apparently functionally isomorphic can invoke very different intuitions.

The third reply I want to discuss is the "other minds reply", which is where Searle's response is perhaps weakest. Harnad (1991) shows that Searle's argument here depends on a "teleportability" to overcome the 'other minds' barrier; a privileged access to Searle-in-the-room. Hauser (1993) takes issue with both Searle's and Harnad's acceptance of the distinction between original and derived intentionality, and describes the problems with this 'other minds' barrier and the "regress to the Cartesian vantage point" (Dennett, 1987) that result. Searle's unfortunate distinction between 'real' and 'as-if' intentionality rules out precisely that privileged access that he has to Searle-in-the-room. This is why the systems reply is so attractive to supporters of artificial intelligence, who almost invariably deny this distinction (Dennett, 1987). The other minds reply is perhaps only a more sophisticated version of the systems reply, but it is more informative because it explains where the understanding lies in the system rather better. Most importantly, it suggests that a lot lies outside the room itself—and therefore outside Searle-in-the-room and the rules and patterns—in the Chinese language itself, for instance, and in the relationship of the room with the world. This reply is especially important to this thesis because the 'other minds' problem is essentially a psychological one, and common-sense psychology is exactly that faculty that we use to deal with this problem every day (Clark, 1987). Searle's "cavalier dismissal" (Hauser, 1993) of this argument—and implicitly of the 'other minds' problem as a whole—isn't justified at all.

Hauser (1993) also points out that Searle indirectly shows that Turing's test *cannot* be taken as a definition, because it is logically possible to pass the test without being able to think; but even so, the test can stand: "the mere logical possibility of a counterexample has absolutely *no* tendency to invalidate a proposed *empirical* test, such as Turing's" (Hauser, 1993, original emphasis). So al-

though Searle has shown the logical possibility of a counterexample, like Block (1981), he hasn't shown the logical or empirical impossibility of an example (Hayes & Ford, 1995), which is what he must do if he is to invalidate strong artificial intelligence.

Next, Searle's thought experiment depends on two features which violate the rules of the Turing test. First, it depends on our knowing the structure of the system under test. It might seem that this shouldn't matter, but to follow Dennett's legal analogy, certain kinds of evidence should probably be declared inadmissible, when this evidence would bias the jury. Secondly, Searle is insisting on our taking the first person point of view, to "identify" (Hofstadter & Dennett, 1981) with Searle-in-the-room, without really taking responsibility for the intuitions that this invokes. On both counts, Searle's experiment actually appeals to those very aspects of common-sense psychology which his approach to the 'other minds' problem ought to deny. On this basis, his thought experiment attacks something like the Turing test, rather than the Turing test itself. Even so, Searle's thought experiment is a very useful tool for probing the intuitions involved in common-sense psychology. For example, Hayes says that cognitive science "consists of a careful and detailed explanation of what's really silly about Searle's Chinese Room argument" (Lucas & Hayes, 1982). Substitute 'intuitive' for 'silly' and you get the point. I'll come back to this in chapters 11, 12, and 13.

The Chinese Room is not just any old philosophical argument. It is, as Dennett (1980) says, an "intuition pump" which Searle uses to explore, in effect, whether or not he can ascribe intentionality (or consciousness) to another system or being. In fact, the Chinese Room is an intuition pump that we can use to study the intuitions involved in ascribing mental properties—and we need to study these intuitions before we can use intuition pumps legitimately. In this thesis, and especially in chapters 11 and 12, I will show that it is possible to build a model of Searle's intuitions and those of some of his opponents in the dialogue surrounding his argument, and to predict, relatively accurately, how these intuitions behave.

This interpretation of the Chinese Room does not attack Searle's philosophical argument (about semantics not being intrinsic in syntax) although there are criticisms which do (e.g. Block, 1986; Boden, 1988). More, it is intended to suggest that detailed and systematic analyses of intuition pumps are a useful methodological tool for studying intuitions about mental phenomena. The

intuitions which are invoked by Searle's intuition pump are those which involve people's common-sense distinction between things which have minds and things which don't. This is why detailed analysis of Searle's intuition pump can be a very useful way to look at what it is to have a mind. After all, philosophers are common-sense psychologists too.

Harnad's Total Turing test

Harnad (1991) proposes the "Total Turing test" (or TTT) which goes beyond the Turing test (or TT) in that it also tests robotic capacities. This is closely aligned with the robot reply to Searle's argument. Harnad's point is that the original intentionality which Searle can't find in the room, could be found if the room was put in the head of a robot with sensors, where there are connections between the world and the inputs to Searle-in-the-robot.

Harnad's response is to put the whole room in the head of a robot and then watch Searle-in-the-robot "seeing", compared to the "understanding" of the original argument. Searle then has two options: either he is accepting symbols from the sensors, in which case Searle-in-the-robot is no longer the only thing in the room, or he is accepting the input to the sensors as well, in which case Searle-in-the-robot is "seeing". Caught in a classic chess fork, the Chinese Room, when grounded by direct sensorimotor action, appears to be a flawed argument. Harnad then argues that the Turing test is implicitly restricted because machines which could pass it don't need to have any ability to "recognise and identify and manipulate and describe real objects, events and states of the affairs in the world" (Harnad, 1991).

Hauser (1993) has a number of criticisms to make on Harnad's modified test, suggesting that his "proposed 'robotic upgrade of the TT to the TTT' is unwarranted". Hauser's main objection is that by accepting, as Harnad does, that ascribed intentionality and real intentionality are categorically different and never the twain shall meet, Harnad is implicitly accepting the point of the Chinese Room argument anyway.

For those who aren't happy even with the strength of the Total Turing test, Harnad (1991) goes one better by offering them the Total Total Turing test, which tests indistinguishability down to the neural and molecular levels. While Harnad himself (Harnad, 1991) thinks that this is unnecessary, the Total Turing test being strong enough, he notes that a few others seem to believe this is the only acceptable version of the Turing test (e.g. Churchland, 1990).

Unfortunately, the Total Total Turing test addresses a completely different question, because it is no longer anything like the Turing test. First off, neuromolecular indistinguishability is not the kind of thing that a human observer can assess. Presumably, the observer is expected to use some artificial aids in this measurement process, in which case the role of the judgement of the observer becomes much more fuzzy. Does the observer trust the machine? If the machine shows that they are different will the observer ever override its indication? Finally, and perhaps most tellingly, which is actually the observer: the human or the machine? A neuromolecular Turing test is operating at the wrong level. It is no longer asking 'can this machine think?' but 'is this machine biologically like a human?' The question is not about mind, but about biology, and therefore it is only a legitimate replacement question under eliminative materialism (Churchland, 1990). Eliminative materialism's position here is at least consistent, but none the more convincing for that.

Finally, it is worth noting that the robot reply argument relies, just like Searle's original argument, on our ability to take the first person point of view. Harnad acknowledges the 'other minds' issues that are involved, but his response breaks many of the same rules as Searle's argument—again appealing to common-sense psychology—and, accordingly, it still needs to be treated with caution.

Moor's inductive interpretation

Moor (1976; 1992) notes that one of the principal reasons for the Turing test being so heavily criticised initially was that it could be taken as having a behaviourist or operationalist basis, as Searle (1980), among others, did. He then points out that the test does not need to be interpreted that way, and proposes an inductive interpretation, in which the test is taken as a format for gathering evidence about whether computers think. He suggests that this is much closer to the 'other

minds' problem; because it is how we deal with other human minds, it should be considered an appropriate way to deal with other computer minds. Dennett's (1985) avoidance of the operationalist's move by interpreting the Turing test as a thought experiment and using a legal analogy has much in common with Moor's treatment of these issues.

This is an excellent way of dealing with some of the more naive and navel-oriented criticisms of the Turing test, and while it neatly escapes many of the attacks on the Turing test, it does attract some brand new ones (Stalker, 1978). The principle, though, seems sound. When studying something elusive which cannot apparently be examined directly, such as a single atom, one of the best ways to do so is to throw things at it and see what happens to them; even though the atom itself is invisible, the patterns that result from these probes will reflect the structure of the atom. Intelligence is just the same (Hofstadter & Dennett, 1981), and the inductive interpretation applies exactly this principle.

An inductive interpretation contrasts really quite sharply with the use of indistinguishability as a scientific criterion (Harnad, 1992). But perhaps the most important point is that this interpretation is dependent on the lifetime of the test being relatively long, and, as in the case of the atom, the kind of probe that is used is critical. It is clear from all versions of the Turing test that the actual questions asked of the system matter critically (French, 1990; Hofstadter & Dennett, 1981). So, to gather information about the common-sense psychology of the system in the test, the probes that are used must be of the right kind to activate this common-sense psychology, whether or not this is compatible with the teletype screen.

A reverse Turing test

In a postscript to "A coffeehouse conversation on the Turing test" Hofstadter (1985) describes a reversed version of the test, where a group of people simulated an artificial intelligence over a teletype link. The event happened at a party, and the fairly predictable result was that the setup was rumbled after an hour or so, but there are a number of informative points in Hofstadter's description.

Perhaps the most telling version of this reversed Turing test was an earlier trial run in which Hofstadter did not participate. In this case, the machine was implemented by a student who "simply had acted himself, without trying to feign mechanicalness in any way". Despite this, none of the observers had seemed to suspect that they were interacting with a human rather than a program. The group was drawn together by Bavel, who afterwards summarised this "by saying that his class was willing to view *anything on a video terminal* as mechanically produced, no matter how sophisticated, insightful, or poetic an utterance it might be" (Hofstadter, 1985, original emphasis). This can be contrasted with Hofstadter's views in the same postscript: "although I don't think it matters for the Turing test in any *fundamental* sense, I do think that which type of 'window' you view another language-using being through has a definite bearing on how *quickly* you can make inferences about that being" (Hofstadter, 1985, original emphasis).

The tension between Hofstadter's view and those of the class is enlightening. First, the mode of interaction does seem to have an effect, and although Hofstadter believes that this effect is merely a delay, I think this conclusion is somewhere between suspect and completely wrong. Second, the Turing test may actually be so hard as to be impossible to pass in practice—even for a human. Of course, an inductive interpretation accepts this, but again it reinforces the absurdity of an operationalist interpretation.

The Turing test as a test for consciousness

An interpretation growing in popularity recently is that consciousness is necessary for much of what we would normally call thinking (e.g. Michie, 1993; Penrose, 1989; Searle, 1990) and, therefore, that either the Turing test is obsolete because it cannot prove consciousness (Michie, 1993) or conversely that it actually is—or should be—a test for consciousness (Penrose, 1989).

Penrose proposes a weaker version of the Turing test in which the "perceptive interrogator should really feel convinced, from the nature of the computer's replies, that there is a *conscious presence* underlying these replies" (Penrose, 1989, original emphasis). Turing (1950) also noted the apparent importance of consciousness to the Turing test in his reply to the "Argument from Consciousness". Most of the time, Penrose seems quite happy with the original formulation of the Turing test, be-

cause of a number of problems that he himself recognises with his proposal (Penrose, 1989). He is, however, trying to make a clear point. He, along with many, if not most, philosophers of mind, is interested in asking the question 'is this machine conscious?' rather than 'can this machine think?'

Flanagan (1993) also points out that although Searle's Chinese Room argument was originally framed as a problem of absent intentionality, it can be interpreted as a problem of absent consciousness. That is, Searle-in-the-room can make the room behave as if it was conscious, while all the time being a rule-following automaton with no consciousness himself. Harnad draws a similar conclusion, pointing out that the other minds barrier normally applies to really knowing whether there's a mind there or not, unless you are in identity with that body, but that "this is exactly what Searle manages to do, because of teleportability" (Harnad, 1991). And Searle himself links understanding to consciousness anyway, arguing that "a capacity for subjective experiences is a necessary condition for having any states with intrinsic intentional content" (Van Gulick, 1988).

Michie's (1993) rejection of the Turing test is based on an operationalist interpretation and an almost pathological solipsism. Unfortunately, he has nothing to replace it with (although he proposes waiting for the "Searle test") and merely accepts the operational interpretation as an interim and scientific variant: "operational awareness". Michie's proposals seem to be too contaminated by operationalism to stay afloat for long—it certainly seems best to avoid mixing operationalism and consciousness in the context of the Turing test.

Taking the Turing test as a test for consciousness is not what Turing wanted, which is why he tried to anticipate it with the "Argument from Consciousness" rebuttal (Turing, 1950), but without total success. While it is clear that the Turing test was not intended as a test for consciousness, the question is: should it have been intended as a test for consciousness? So far, the evidence is that consciousness is probably correlated with what we would recognise as thinking in ourselves, but it is as yet unclear how necessary or sufficient it actually is for thinking itself. The Turing test should not be, and probably can never be, *only* a test for consciousness.

Interim summary: human and computer minds revisited

This plethora of interpretations of the Turing test doesn't do very much to create a unified view of the problems and issues that it raises. There are, however, a number of recurrent themes, and I will review these apparently fundamental asymmetries between the participants in the Turing test, before moving on to address some of the different perspectives involved.

First, and most obviously, human minds and computer minds are physically different; embodied in different kinds of stuff, protein on the one hand, and silicon on the other. One class of criticisms of the test rests on whether this difference matters, but the original Turing test hides these kinds of details, so the observer should not be aware of the difference. A few brave souls have claimed that there are logical constraints on physical substrate (e.g. Searle, 1980; Penrose, 1989) but the lack of a counterexemplary silicon-based mind does not constitute proof (Turing, 1950).

Second, human minds and computer minds are socially different. This may be a criticism of computers today rather than computers *per se*, because projects like Brooks and Stein's COG (Brooks & Stein, 1993) propose a codevelopment of humans and android computers. In this sense, the computer may be a kind of 'child' growing up in a human society, and will therefore be socialised into it (Collins, 1990). Turing (1950) also alludes to this possibility. In other words, Wittgensteinian language-games between humans and computers may be possible, and may help bridge the social gap between them.

Third, and perhaps most worryingly, human and computer minds are evolutionarily different. This is the explicit criticism of Humphrey (1992) and can be seen implicitly in Searle (1992), too. Humans have evolved over billions of years, and have carried traits through that history. No doubt many of these traits are irrelevant to us today, but it is not possible to tell which, because the original meaning has often been lost. Criticisms in this class seem to carry most weight at the moment, but because they are essentially unproven—and apparently unprovable—artificial intelligence can acknowledge them as a potential problem and carry on a wary research plan—tentatively avoiding the obviously evolutionarily incorrect paths.

None of these differences between human and computer minds precludes the possibility of artificial intelligence in general, or computer implementation of common-sense psychology in particular, although they do make it rather difficult. But as they are fundamental asymmetries between the participants in the Turing test, in order to make more sense of the test even as a thought experiment, it is important to examine the roles of the factors which influence the outcome of the Turing test. I will discuss these factors in the following sections.

On the role of the observer

The observer's role in a Turing test is not quite as fixed as it might have seemed in Turing's original article. The psychological baggage of the observer plays a key role in the many of the interpretations, implicitly if not explicitly (e.g. Collins, 1990; Harnad, 1991; Hofstadter & Dennett, 1981; Searle, 1980). This dependence on the psychology of the observer means that the test cannot be taken as an operational or behaviourist test as it stands.

The idea is simple: the observer's knowledge and expertise fundamentally influences the kinds of judgement that they will make in a Turing test. Later, in chapter 13, I will argue that this is because the Turing test actually exercises the observer's common-sense psychology at least as much as the system's intelligent behaviour, and for this reason, a variation on the test which looks principally at the observer's ascription of mental states may offer useful insights into the nature of the human common-sense psychology. In practice, an observer's ascription of mental states to something seems to be dependent on their expertise, prejudices, and understanding of cognitive science, among other factors, and is therefore subject to all sorts of subtle and individual variations.

The connection between the Turing test and the philosophical 'other minds' problem has been noted by both Harnad and Hauser in their exchange, and despite Searle's attempts to brush it under the carpet there does seem to be a general problem of deciding whether another body has a mind, independent of the actual form of that body. The question is, what kind of decision is involved in ascribing a mind in the Turing test: is it a scientific one or a common-sense psychological one?

Depending on the scientific or common-sense role of the observer, the Turing test can vary between a behaviourist or operational (but meaningless) definition of thinking and a test for interactional compatibility between two different systems. These alternatives raise very different issues.

On the role of bodily appearance

Turing's original test rules out bodily appearance as prejudicial to the integrity of the test, and uses a teletype link to screen it out. Harnad (1991) comments: "bodily appearance clearly matters far less now than it did in Turing's day: our intuitive judgement about an otherwise convincing candidate no longer runs much risk of being biased by a robotic exterior in the Star-Wars era, with its lovable tin heroes". But robots are nothing new; the Tin Man in the film of *The Wizard of Oz* dates back to 1939, more than ten years before Turing's article was published. In fact, it is far from clear that bodily appearance is becoming less relevant; these fictional robots are all to some extent anthropomorphic, and have bodily appearances and behaviours that approximate ours to a greater or lesser degree. And this effect of anthropomorphism certainly seems to increase the tendency we have to identify with robots in general.

Anthropomorphism is a natural phenomenon: we do it all the time, often without even noticing (Caporael, 1986; Krementsov & Todes, 1991). We can anthropomorphise things like rocks, which have no semblance of human form, but we do seem to find it easier if they approximate us physically. Our first judgement would be biased by a human form, but it would be a bias in favour of the system *passing* the test, rather than failing it. Psychological studies show that teleological reasoning and mental states tend to be attributed to animals not uniformly, but in some sense according to how close they are to humans (Eddy *et al.*, 1993; Tamir & Zohar, 1991). The role of anthropomorphism in this ascription of cognitive powers is a fundamental theme in this thesis; one which I will return to in much more detail later in chapter 7.

Bodily appearance, to some extent at least, does seem to be important in forming judgements. Technology has changed such that a virtual reality version of the Turing test is not so far away, and probably Turing would have suggested it with today's technology. So while we may agree with Harnad that robotic exteriors might well be more acceptable than they used to be, it is still impor-

tant to recognise that the effect of bodily appearance hasn't gone away. If the bodily appearance is visible, it will have an effect on the Turing test. But this leaves another question: what happens to the ascription if we try to hide bodily appearance?

On the role of the modality of interaction

Turing's restriction of the interaction to a purely linguistic exchange between the system and the observer seems at first glance to be innocuous, but the tendency of a teletype version of the Turing test to bias the observer into sometimes guessing incorrectly that the system is a machine, despite all the available evidence, was shown by the reverse Turing test trick played on Hofstadter (Hofstadter, 1985).

In another experiment, closely related to the Turing test, Garfinkel (1967) gave students a counselling system which required all questions to be framed such they could only take 'yes' or 'no' answers. The students thought they were interacting with a human counsellor, but in practice all questions were answered randomly, as if by tossing a coin. The students were quick to see significances in the answers, and even when the answer contradicted a previous one "the underlying pattern was elaborated and compounded over the series of exchanges and was accommodated to each present 'answer' so as to maintain the 'course of advice'" (Garfinkel, 1967). The underlying mechanism is even more trivial than that of ELIZA (Weizenbaum, 1966) and PARRY (Colby, 1981), and it still managed to 'fool' people. Of course, the students weren't expecting this kind of a system and therefore weren't forming critical judgements, but even so it is as if the restricted interaction modality and the restricted mechanism somehow cancel each other out, still allowing the observer to ascribe the full gamut of mental properties.

These experiments show that the screen Turing introduced does a lot more than hide the bodily appearance: it changes the patterns of interaction involved. In a sense, it replaces the dependency on bodily appearance with a different but closely related dependency, which plays a very similar role in the ascription of mental states. It is as if the reduced interaction medium isn't hiding the bodily appearance, it is just creating a different one, so all the effects of bodily appearance are still there, but they have been transferred to a different object.

On the role of the knowledge of the mechanism

As Searle accidentally showed in his Chinese Room thought experiment, knowledge about what is going on inside a system does affect the test. This is what Turing was trying to avoid by specifying that the bodily appearance was hidden, but that having been overtaken by the Harnad's Star-Wars effect, we are left with the problem that as soon as we know how a system's internal mechanism works we are more likely to see it failing the test rather than passing it. Eccles (1964) makes this point explicitly.

This effect is often most intense with the consciousness variation of the Turing test, because it is the knowledge of the mechanism that seems to prevent us from accepting that there is something it is like to be the system under test (Nagel, 1974). It is Searle's insistence on the first person point of view that makes us identify—or, as in Searle's case, *fail* to identify (Searle, 1980; Searle, 1992)—with the system. To the extent that this identification is a result of the observer's common-sense psychology this is an important effect, and one which I will analyse in more detail in chapters 11 and 12.

Turing's point behind the problematic teletype connection was to hide the actual form of the system from the observer, to mitigate prejudices about the mechanism, but as I have already discussed, the effects of the screen are more far-reaching than this. The teletype screen Turing proposed is far from passive, and the interaction that results shows a significantly different pattern of judgements from that which would be expected of face-to-face interaction (Hofstadter & Dennett, 1981).

On the role of the social context

An aspect of Turing's paper which is rarely referred to is the point he makes about the test being part of a changing social context: "I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted" (Turing, 1950). Perhaps our use of words is changing, but it doesn't seem to have changed as fast as Turing thought it would—or perhaps there has been

a divergence between the common-sense use of the word 'think' and the philosophical use of 'think' which Turing didn't anticipate. As far as common sense is concerned, we do talk about machines, even computers, 'thinking' all the time, although if we switch into a different context, 'philosopher mode', we see them entirely differently. The way that we choose to see the system, the "stance" (Dennett, 1971) that we take, critically affects the judgements that we make.

Perhaps this isn't a new phenomenon. Darwin (1871) describes the conflict over whether the different races of the human species should be classified as different species or as variation within one. Darwin's conclusion was the one that would be universal today, that there is a single species, but he cites other commentators who suggested dividing it into between two and sixty-three distinct species. Much hinges on the precise notion of a species, and racism is still widespread, but we now live in a world where animals can be said to have rights, and where mammals at least are usually taken as being to some extent conscious. In biology it seems that the gap between species is closing as knowledge increases. This, it would seem, was the kind of trend that Turing expected to continue. The phrase 'can machines think?' should not be taken as constant, because 'machine' and 'thinking' are words that can change their meaning as our society gradually becomes a symbiosis of different technologies, and the modal force of 'can' allows for this future revision.

Summary: common-sense psychology in the Turing test

So what of the Turing test? There are a number of issues it raises. Whatever Turing's original intention, it shouldn't be taken as an operationalist definition, because this provides no insight into the problem of real intelligent behaviour, and also because it leaves the test impossible to standardise. In fact, it is unclear that the Turing test provides any scientific insight at all into the problems of real intelligent behaviour; that is, the operationalist interpretation stretches the test beyond its breaking point. The relativistic nature of the test means that taking Turing indistinguishability as a scientific criterion is probably not appropriate, despite Harnad's arguments on this point (Harnad, 1992).

But taking the Turing test as a thought experiment, perhaps through an inductive interpretation (Moor, 1976), does offer some enlightening insights. Between them, the different analyses show that the test depends on several factors, one of which is the behaviour of the system, but equally important are the physical appearance of the system, the interaction modality, social context, and the psychology of the observer. As a simple format for evaluating these factors and their interactions, the simplicity of the Turing test has much to be said for it. I will return to these factors in chapters 7 and 8, and show how they give important clues to the structure of common-sense psychology.

To summarise, the Turing test is flawed if interpreted in any sense as a scientific test for intelligence, but it has real potential as a tool for investigating some aspects of intelligence, particularly the common-sense psychology of the observer. This inversion, where the test is used to investigate the observer as much as, if not more than, the system under test, is a key aspect of this thesis, and a more detailed analysis of it will be presented in chapter 13.

In these review chapters, I have shown that common-sense psychology is important to philosophy, psychology, and artificial intelligence. In the next parts of the thesis, I will argue the case for its central importance to artificial intelligence more fully, construct a descriptive theory and model of common-sense psychology, and use it to analyse two studies of the ascription of mental states. These two studies, Baron-Cohen *et al.*'s false belief test, described in chapter 3, and Searle's Chinese Room thought experiment, described earlier in this chapter, show that common-sense psychology affects cognitive science in relatively predictable ways, and that a better understanding of the effects of common-sense psychology can lay down foundations for a better future in cognitive science.

Part Two

Anthropomorphism in common-sense psychology

BLANK IN ORIGINAL

Chapter 6

A common-sense psychology manifesto

Introduction: from naive physics to naive psychology

More than ten years ago, Hayes published the "Second naive physics manifesto" (Hayes, 1985b) as a critique of contemporary research in artificial intelligence. Hayes proposed that we "put away childish things by building large-scale formalisations", beginning with "our knowledge of the everyday physical world". He, and others, have since put a lot of effort into developing models of our common-sense understanding of the physical world.

But life in the common-sense community isn't a bed of roses. Common sense has generally been a big problem for artificial intelligence, and despite the attempts of many brave souls (e.g. Hayes, McCarthy, McDermott, and Lenat) it hasn't really yielded substantially: "the commonsense knowledge problem has blocked all progress in theoretical artificial intelligence" (Dreyfus & Dreyfus, 1988). For twenty years there has been only slow progress in the field, and although many important technologies have been found while looking for ways to deal with common sense (e.g. nonmonotonic reasoning), the core of the problem is still almost untouched.

This chapter argues that artificial intelligence has failed to address the whole problem of common sense, and that this is the cause of the recent stagnation in the field. Perhaps the biggest gap is in common-sense—or naive—psychology, our natural human ability to see one another as minds rather than as bodies. This is especially important to artificial intelligence because, among other things, artificial intelligence must eventually enable us humans to see computers not as grey boxes, but as *minds*. In this chapter, I will argue that we should study exactly this—what is going on in people's heads that makes them see others as having minds.

A revised version of this chapter has been published as "A Brief Naive Psychology Manifesto" in *Informatica, 19*(4), 1995, pp. 495-500, a special issue on the differences between minds and computers.

The problems for common sense have generally come in two classes, technical and methodological. On the technical side, for example, there is the 'frame problem' (McCarthy & Hayes, 1969) discussed in chapter 4, which Dennett (1984) describes as the "smoking pistol" behind a lot of the attacks on artificial intelligence, and as a major philosophical problem for everybody. I have already reviewed some of the methodological problems for the different approaches to common-sense reasoning in chapter 4, but, in general, these methodological problems only affect research directed at producing causal accounts of common-sense reasoning, not descriptive ones (Clark, 1988; McDermott, 1987).

This chapter will present a position orthogonal to that of Hayes. Hayes' initiative clarified many of the issues associated with common sense, and other developments in comparative and developmental psychology have further highlighted the apparently fundamental nature of common-sense physics (Carey, 1985; Premack & Woodruff, 1978), but they have also revealed a deeper and bigger problem than that of common-sense physics—common-sense psychology. An important feature of common-sense psychology is that it seems to be responsible for the faculty that normal adult people have which enables them, in short, to see one another as minds rather than bodies. This is an issue that artificial intelligence also needs to address. While people see one another as minds not simply bodies, they just don't see computers as minds in the same way (Caporael, 1986; Turkle, 1984). If artificial intelligence is to overcome this barrier, we humans must be able to see minds in artifacts, to ascribe mental states to artificial intelligences in the same way that we do to people. Unfortunately, this is partly a property of people, not purely of artificial intelligence systems, so it has mostly been overlooked in artificial intelligence research.

An example from Haugeland (1978) will make this clearer. Imagine that somebody presents you with an object and tells you that it can play chess. How can they convince you that it really does? You can look at the behaviour of the system, but that might be open to different interpretations depending on how the system communicates its moves. You decide by looking at its behaviour, interpreting it in a consistent way, and checking that the moves that it communicates are legal and sensible with respect to the rules of chess as you know them. Haugeland's point is that the mean-

ing of the behaviour isn't an intrinsic property of the behaviour, but is "attributed in an empirically justified interpretation" (Haugeland, 1978). But this ability to make the interpretation is itself in part a property of you, the judge.

Now expand the idea to the whole of human interaction, which is similarly game-like. The 'moves' can again be interpreted in many different ways, but if the moves are to tell a coherent story, they must be compatible with the laws and heuristics of other people's interpretation. Human interaction, like chess, isn't usually learned from books of rules, but by people learning from one another and from experience. It depends more on human agreement than adherence to sets of explicitly specified rules. Just as learning chess from watching only one player's moves would be almost impossible, the same is true of human social interaction. And any good chess program will depend on a player's *understanding* of the moves, in terms of (often social) metaphors like 'attacking' and 'supporting'.

Following on from this game metaphor for human interaction, as I discussed in chapter 3 there is evidence that a lot of human intelligence is 'Machiavellian' (Byrne & Whiten, 1988) in the way people use it to outwit each other and to recognise and manipulate one another's mental states. Our social environments are considerably more complex than our physical ones, and survival in these social environments required us to develop common-sense psychology so we could recognise and reason about one another's mental states from observing their behaviour (Humphrey, 1976). Humphrey even suggests that common-sense physics itself may in part be derived from a leaky common-sense psychology, and that it is common-sense psychology that is at the core of our understanding of the world; he suggests that the transactional nature of human social interaction is so fundamental to our nature that even in situations involving physical actions, we sometimes see these actions as if they are *trans*actions.

At the heart of this proposal is a methodological inversion. Usually, artificial intelligence is thought of as 'the science of smart behaviour'—building systems which behave in a way that seems 'intelligent'. This naturally leads to all sorts of definitions and operational interpretations of the word 'intelligent' none of which really help to find intelligence. The reason they don't help is they miss the point: a definition of intelligence has to become part of science, but doesn't necessarily have

any impact where it counts, which is on everyday human common-sense psychology. This is because the word 'intelligent' doesn't refer to a scientific property, but to an everyday human judgement.

So this is only part of the story. Artificial intelligence also needs to study common-sense psychology to find out what is going on in my head to make me see other people as having minds—and it is this that is an inversion of the conventional artificial intelligence approach. There should be two parts to the study of intelligence: the research on smart behaviour that we're all familiar with, but research is also needed on our ability to *recognise* that behaviour as smart. The sin of artificial intelligence is a sin of omission—it hasn't completely addressed the second part of the problem, that of common-sense psychology. Common-sense psychology is not more important or significant than other abilities, but it is an equally important element of human cognition, and, further, it is an essential part of how we recognise intelligence. It should, therefore, be a topic for serious research in artificial intelligence.

A brief history of common-sense psychology

Common-sense psychology isn't new to artificial intelligence, which has already tried a number of approaches to the problem. Perhaps the most successful have been the axiomatic formalisms reviewed in chapter 4 (e.g. Cohen & Levesque, 1990; Moore, 1985). These represent common-sense psychology as the ability to make inferences about a set of beliefs, desires, and intentions, corresponding to an agent's mental states. These axiomatic formal approaches to common-sense psychology are usually based on some kind of modal logic, with knowledge and belief, desires and intentions as different operators in that logic. These logics enable representation and reasoning about someone's mental states by making these states unobservable to others, so these 'opaque' agents can, for example, believe something which other people know to be false.

I have already discussed the problems with most approaches to common-sense psychology in artificial intelligence in chapter 4. Generally speaking, there are both technical and methodological criticisms. The methodological criticisms are the less serious ones; for the most part they simply argue that the kinds of representation used are inappropriate in practice, usually because

they leave something out. McCarthy and Hayes' (1969) methodology, for instance, leaves any aspects of their representations' "heuristic adequacy"—how the representations are actually used when solving problems—to future research.

The theoretical problems of common-sense psychology in artificial intelligence are a bit more significant. Some of these problems are specific to common-sense psychology, for example, as I discussed in chapters 3 and 4, it is far from clear that the coarse classification of common-sense mental states into the attitudes of beliefs, desires, and intentions is valid (Dennett, 1987; Samet, 1993). This is an assumption, and while it works well a lot of the time, there are some mental states which fit uneasily in this model (e.g. moods and hostility). A lot of these are dispositional in the manner of Ryle (1949), who shows that a considerable part of typical human conversation is not in the indicative mood of propositional statements.

Secondly, there are more general criticisms of representational artificial intelligence which apply only indirectly to common-sense psychology, for example, the situated (e.g. Brooks, 1991b) and connectionist (e.g. McLelland, Rumelhart, & Hinton, 1986) paradigms seem to deal with some problems which are virtually insoluble in traditional representational artificial intelligence. Even so, as I argued in chapter 4, they haven't much impact on common-sense psychology, because they have to deal with something rare in physical environments, this opacity that people have. Both approaches break down with the opaque nature of other agents. While they are good at dealing with the observables of the physical world, they are less good at dealing with the unobservables of other agents' mental states. This really does seem to be what representations are good at. If representations have to become situated to deal with the physical world, situated approaches have to become representational to deal with the psychological one.

So there has been significant work in this field, but perhaps artificial intelligence just hasn't realised the scale or importance of the problem. Common-sense psychology is *the* pressing problem. All the major philosophical stumbling blocks of artificial intelligence, including consciousness and intentionality, can be traced partly to our inability to understand when and how to ascribe mental states to computers or other artifacts. This doesn't mean that common-sense psychology is logically prior to these problems, but that it is *methodologically* prior. If artificial intelligence is to

tackle consciousness and subjectivity, it must also (and first) study the complex of intuitions involved in our *recognition* of consciousness and subjectivity—it is not enough to play operationalist games with it (Michie, 1993).

This proposal is not one of "psychologising", making "the artificial imitate the natural" (Israel, 1985), although even this might turn out to be necessary after more research in the area. The point is that we—the people who are designing and evaluating these machines—are people with relatively uniform cultures, societies, and biologies—at least when compared to machines. Perhaps, as Searle (1992) claims, these factors affect human mental phenomena directly. But even if they don't, they may well affect our recognition and interpretation of those phenomena, and therefore, our ability to ascribe mental states to machines.

We could, of course, take McCarthy's stance: "this is artificial intelligence and so we don't care if it's psychologically real" (Kolata, 1982). But as soon as we talk about minds we are, whether we like it or not, involved in something psychological, so to compare minds and computers will inevitably be a partly psychological question. The actual nature of the distinction between human intelligence and artificial intelligence cannot be ignored.

Human and alien intelligence

Within artificial intelligence, there is often an assumption that there is something which can be called 'intelligence', but which is more abstract or general than what we call 'intelligence' in people. We can call this the 'alien intelligence hypothesis'. It is entirely possible that the alien intelligence hypothesis is simply false. If the complex bag of phenomena we call 'intelligence' is something that people use to interact with each other in human societies, an alien intelligence which doesn't interact in the same way in these societies might not be seen by us as intelligence at all. Perhaps systems are just seen as intelligent in proportion to how well we can understand their patterns of behaviour, as Haugeland's chess playing machine example shows.

There are three possibilities for artificial intelligence. First, it might be possible to build a system with 'alien intelligence' which doesn't require our human common-sense psychology to see it as intelligent. Second, it is possible that 'intelligence' implicitly means 'human intelligence'—that is, it is the same kind of thing that we recognise in each other—and therefore it appeals to our human common-sense psychology to recognise it as such. A third possibility is that intelligent systems can never be built (either for logical or pragmatic reasons) in which case common-sense psychology isn't relevant. Dealing fully with this third possibility is outside the scope of this thesis; it will suffice for now to say that there are no clearly sound arguments for the logical impossibility of artificial intelligence, although there have been a number of brave attempts (Penrose, 1994; Searle, 1980).

A simple reply to the first possibility would be that supposing alien intelligence did exist, how would we recognise it as such *without* appealing to our human recognition of intelligence? The idea that alien intelligence exists but isn't recognisable is unfalsifiable. We could call this the 'Vegetarian's Dilemma', lettuce might be conscious and suffering immensely when we eat it, but if it is incapable of communicating that fact to us, should we ignore the problem? On this principle, computers (along with lettuce, beer cans, and rocks) could already be intelligent, we just aren't able to recognise them as such. If intelligence is recognisable, it must be recognisable by us humans, and then it probably doesn't count as alien intelligence any more.

This is partly because human intelligence isn't restricted to any particular narrow set of behaviours. What we recognise as intelligence can change—indeed it is pretty certain that computers have already introduced such a change. One of Turing's (1950) points was that much hangs on "the use of words and general educated opinion". Artificial intelligence isn't just a case of making machines smarter, but also of making us better at seeing their smartness.

Perhaps the case for the alien intelligence hypothesis is poorer than it seemed at first. The onus is on those who claim that systems have alien intelligence to *demonstrate*, not simply define, its independence from human intelligence. If there is no such thing as "intelligence in general" (French, 1990)—if the principles of thinking for artificial intelligence are the principles of thinking for human intelligence—we are reduced to the second possibility, and should take the term 'intelli-

gence' to mean culturally-oriented human intelligence, as Collins (1990) does. After all, this is what the Turing test really tests for (French, 1990), which is perhaps why its results are sometimes so strange (Colby, 1981; Collins, 1990; Hofstadter & Dennett, 1981; Weizenbaum, 1976). As I said in chapter 5, the Turing test is, in many ways, a paradigm example of the effects of common-sense psychology (Caporael, 1986), in that it is, in effect, a test of the ability of one system to ascribe mental states to another, not of any objective behaviour of any single system. Most of the interesting stuff in the Turing test is going on in the head of the observer, not just in the system under test (Collins, 1990). The Turing test is not, as Searle (1992) takes it to be, "third person or 'objective'". We might even suggest an inverted Turing test: putting a system in the role of the observer and looking at its ability to distinguish between humans and programs could be a useful way of evaluating its common-sense psychology. I will expand on this idea in chapter 13, and argue the case for this inverted test more fully then.

The problem is this persistent anthropocentricity—we can't step outside our humanity although we perpetually see things as if they are in some way independent of our humanity. While for physics that doesn't seem to matter much, for psychological concepts such as 'intelligence', we must remember that we are human. We need to discover what it is like to be human before we can truly know what the differences between people and machines are, and this is hard for us humans!

Models for common-sense psychology

In looking at what is going on in our heads when we see people as minds rather than as bodies, some of the most useful tools are models of the process of ascribing mental states to other systems. In this section, three candidate models will be examined in a little more detail, anthropomorphism, the simulation model, and the theory model.

Model one: anthropomorphism

One way of ascribing mental states to a system is just to anthropomorphise it—to ascribe it human mental characteristics without reference to its real competences. Anthropomorphism is a complex and subtle phenomenon (Caporael, 1986; Eddy *et al.*, 1993) and not one that has been studied much. It is, though, at the heart of this thesis, so I will discuss it in considerably more detail in the next two chapters.

When Eddy *et al.* (1993) studied people's psychological tendency to anthropomorphise animals, they found that there are two primary mechanisms involved: "people are likely to attribute similar experiences and cognitive abilities to other animals based on (1) the degree of physical similarity between themselves and the species in question (e.g. primates), and (2) the degree to which they have formed an attachment bond with a particular animal (e.g. dogs and cats)" (Eddy *et al.*, 1993). Computers don't score too well on physical similarity, so this is likely to form a persistent bias against people ascribing mental states to them, unless we build them with a physical resemblance to us (e.g. Brooks & Stein, 1993). Familiarity, fortunately, offers us a way out of this trap—we can in principle learn to see computers as minds, for example, while carrying out a Turing test.

There are several possible theories of anthropomorphism. Caporael (1986) suggests that it is a "'default schema' applied to non-social objects, one that is abandoned or modified in the face of contradictory evidence", but the evidence is against either animals or computers really being 'non-social' (Haraway, 1992; Nass, Steuer, & Tauber, 1994) and familiarity can increase rather than decrease the tendency to anthropomorphise (Eddy *et al.*, 1993). Alternatively, perhaps, our tendency to anthropomorphise is really a disposition to take the "intentional stance" (Dennett, 1971), to see others as minds rather than as bodies. If, instead of taking the intentional stance, the physical stance is taken, the very different faculty of common-sense physics will be deployed. Anthropomorphism, then, could be a part of common-sense psychology in that it determines whether or not an intentional stance will be taken, but it is not truly part of the stance itself. Instead, it plays the role of the rationality assumption in Dennett's model—although clearly an-

thropomorphism isn't the same thing as rationality. The suggestion that the rationality assumption is, in Dennett's (1971) words, "pre-theoretic", does perhaps give us the scope to interpret it this way.

Model two: simulation

As I discussed in chapter 3, sometimes one person's prediction of another person's mental states can be thought of as 'simulating' the other person, pretending to be them, and looking at the world from their point of view. Clark (1988) suggests that a similar simulation process could even account for common-sense physics—perhaps Hayes' (1985a) representation of the behaviour of liquids could be recast as a kind of simulation, and as far as the predictions are concerned, viewed externally, there needn't be any difference. For common-sense psychology, there is evidence that for some predictions—particularly those involving affective states—an ability to simulate other people works well (Hobson, 1993; Perner, 1991). And representational artificial intelligence does simulation all the time—as look ahead in game playing for example. Simulation, or taking another person's point of view, is a way that we can understand some aspects of another's mental states; for instance, to recognise somebody's ignorance (Davis, 1988).

So simulation is another way that we can reason about other people's mental states. It is a way that works rather better for affective than for cognitive states (Hobson, 1993) but it doesn't deal with everything: there are some tasks which children actually answer differently, but which they ought to answer the same if they use simulation to get the answer (Perner, 1994). Something is left over, and that something is a 'theory' of mind—not a theory in the scientific sense (Clark, 1987; Searle, 1992)—simply a theory in the sense of a set of tools for thinking about the unobservables of another person's mental states.

Model three: theory

So finally, there is the theory aspect of prediction discussed in chapter 3; it is that aspect which is most similar to representational artificial intelligence. Some (e.g. Fodor, 1985) even take it as the complete answer to common-sense psychology, but this stretches it too far; a strong representational theory of mind is subject to too many philosophical and evolutionary objections, as I discussed in chapter 2, and besides, it fails to account for all phenomena (Hobson, 1993; Perner, 1994). But just because a representational theory of mind can't provide a complete common-sense psychology that doesn't mean that it can't form part of a complete common-sense psychology. The theory theory, as it is currently interpreted in psychology, describes common-sense psychology as a set of rules and tricks for dealing with the unobservable mental states of others. Its best analogue in artificial intelligence, therefore, would be a body of laws and heuristics for guessing at other people's mental states. Young children, for example, can't properly ascribe beliefs they know to be false to someone. Their ascription might be described by a rule like:

$$\textit{believes(self, } X\textit{)} \Rightarrow$$
$$\textit{believes(self, believes(agent, } X\textit{))}$$

That is, I believe that *agent* believes X if I believe X. In time, the patterns of ascription change, so that the child becomes aware of the other person's perceptions. So now the ascription can be described:

$$\textit{believes(self, perceived(agent, } X\textit{))} \wedge \neg\textit{believes(self, believes(agent, } \neg X\textit{))} \Rightarrow$$
$$\textit{believes(self, believes(agent, } X\textit{))}$$

That is, I believe that *agent* believes X if I saw *agent* perceive X, and I don't believe that *agent* believes X is false. These rules are steps towards a simple descriptive account of how an agent might come to pass the false belief test (Baron-Cohen *et al.*, 1985). Note that both of these rules are intentional, that is, they only apply when we see *agent* as being something we can reason about psychologically rather than physically; in Dennett's (1987) terms, this means they only apply when we take the intentional stance to *agent*.

Whether this account is correct or not isn't really the issue. These rules, which have a close affinity to those of Davis (1988), are a first step to modelling Wellman's (1990) account of belief progressing from a copy theory to a representation theory, but there are many other possible models (e.g. Baron-Cohen *et al.*, 1985; Perner, 1991). What is actually needed is a way of comparing and combining the different models, and this is where artificial intelligence can help (Samet, 1993).

In artificial intelligence, all the best programs for playing games like chess (and games are often a good metaphor for human social interaction) use a subtle mixture of simulation (look ahead) and theory (heuristics) because neither on its own is ever sufficient. In principle, of course, a heuristic theory can generate a simulation (Davies, 1994), and in practice an actual system—such as a trained connectionist network—might show aspects of theory and simulation under different circumstances, just as electrons can behave like particles or like waves, depending on the experiment and the environment. These are valid points, but perhaps they are more relevant to a causal account than to a descriptive one—and a causal account is beyond the scope of this thesis.

These three models—anthropomorphism, simulation, and theory—represent different aspects of common-sense psychology rather than the whole, but they can be combined to create a composite model. When trying to predict or reason about the behaviour of a system, a complex of dispositions, one of which is anthropomorphism, selects a stance with respect to that system. These stances deploy natural faculties—so when dealing with a physical system, common-sense physics is applied, but for a psychological system, common-sense psychology is applied. Often, both stances can, in principle, be taken to the same system at the same time (even a thermostat, McCarthy, 1979), although in practice there seems to be a mutual exclusion between the different stances (Carey, 1985; Dennett, 1971). Through anthropomorphism individual differences in the dispositions and the social context can influence how different people can see the same system differently.

A mind will only be seen in the system from the intentional stance (Dennett, 1971)—that selected by anthropomorphism—and within the intentional stance as a whole there may even be different sub-stances which depend on the access that is required to the other's mental states. If we are to

'simulate' it—to see what it is like to *be* the system—that may only happen if the system is believed to have the right kind of mechanism. The theory stance, on the other hand, may be better at dealing with external, behavioural, questions.

So how well do these models do? Although they only hint at the true complexity of common-sense psychology, they do have some predictive power—this can be shown with Searle's (1980) Chinese Room thought experiment. I have already discussed Searle's argument in a little detail in the previous chapter, and this is not yet the place for a new rebuttal or any fuller analysis of Searle's argument. Here I will only claim that, as I implied earlier, intuition plays an important role in Searle's thought experiment, and the intuitive part of the thought experiment is attractive because Searle insists on us taking the first person stance when we look at the scenario. This is a big hint that he is appealing to our common-sense psychology as we try to understand his model.

First, Searle claims the system can be assumed to have passed the Turing test, so as far as external appearances are concerned, it has already been ascribed mental states. It will probably be difficult to anthropomorphise a real room on the similarity factor, but the Turing test offers time over which to learn to ascribe a mind to the room, given the fact that its responses are indistinguishable from those of a person. But then Searle adds something—he tells us what is going on *inside* the room; in fact, he tells us that inside there is an agent, Searle-in-the-room, manipulating slips and scripts according to a rule book written in English. Now the story is different: we can anthropomorphise either the room, Searle-in-the-room or the rules and scripts, but Searle-in-the-room (being human) is by far the most similar to us, so he acts like a magnet to our common-sense psychology and we are pushed into taking his point of view—which is forbidden by Searle's rules. We are then trapped, as Hofstadter was, asking "which level does Searle wish us to identify with?" (Hofstadter & Dennett, 1981). And variations on the theme may change the identification: the story changes if we put the room in the head of a robot, partly because the head of a robot is a lot easier to identify with than a room.

This point mustn't be pushed too far yet, but it shows how sensitive thought experiments can be to common-sense psychology, and that even a fairly simple model can begin to make rough predictions about intuitions generated by common-sense psychology—even a philosopher's com-

mon-sense psychology. We can learn a lot by developing a model which matches Searle's intuitions, and, accommodating individual differences, those of his critics; I will develop just such a model later in chapters 11 and 12, and show how central common-sense psychology is to these variations in different people's intuitions. The idea of a computational model of a philosopher is perhaps rather bewildering, but if we borrow from McCarthy the idea that a part of philosophy is "the science of common sense" (McCarthy, 1979) perhaps a computational version of this part isn't so strange after all.

Another example of the effects of common-sense psychology is Woolgar's (1985) description of a device which bolts on to a video recorder and splices out advertisements during recording. On one level this is intelligent behaviour, but if you then read the instructions, and they tell you that the device actually works by detecting a particular signal in the transmission, this changes the ascription of intelligence, and "redefines and thus reserves the attribute of 'intelligence' for some future assessment of performance" (Woolgar, 1985). Again, the change in our knowledge affects the stance that we take—affects whether or not we see the system from the intentional stance.

This deeper study of the ascription of intelligence shows a sensitivity to physical form and our knowledge of the system's design which is perhaps rather distressing for strong artificial intelligence. It seems to show not that artificial intelligence is impossible in principle, just that it can be very hard for people to see things which don't physically, structurally, and behaviourally resemble humans as being intelligent. Perhaps Brooks and Stein (1993) were right to design COG with a humanoid form, not for any technical reason, but simply because it will make it easier for us to see COG as an intelligent system.

Common-sense psychology for machines

Hayes (1985b) concluded the "Second naive physics manifesto" with a discussion of the importance of common sense for artificial intelligence. While I would agree that common sense is too important to fall with the problems of McCarthy and Hayes' methodology, the motivations for this proposal of serious research on common-sense psychology are rather different. Of all the common-sense disciplines proposed by Hayes and others, common-sense psychology is the only

one that is obviously specifically human, but in all this work there is an implicit anthropocentricity. Right back to McCarthy's (1959) proposal of common sense, it was assumed that the common sense to be used was human common sense.

As I've said, it is entirely possible that intelligent behaviour is distinguished not by an objective criterion of success, rationality, adaptiveness, or what have you, but by a more subjective criterion of compatibility with our human common-sense psychology. If this is true, there are two ways to build intelligent systems. The first is to build such systems, and then to look at them to see whether we think they are intelligent or not, and to use the insights we gain to change our design approaches for the future. A second strategy is to look first at how human common-sense psychology construes more familiar systems. The problems of the first strategy are simply that, without any principled understanding of how we construe the behaviour of such systems, we stumble across all sorts of psychological biases. These biases can make us completely misinterpret the behaviour of systems, so a better understanding of the biases themselves is needed before we can be certain about our understanding of the behaviour of systems.

There is no strong methodological component to this proposal, largely because the project is just too important to be dismissed purely on methodological grounds—and the same point goes for common-sense physics. Dreyfus and Dreyfus claim, for example, that "the problem of finding a *theory* of common-sense physics is insoluble because the domain has no theoretical structure" (Dreyfus & Dreyfus, 1988). Well, this depends on what you want from a theory. Even if, as Dreyfus and Dreyfus claim, common-sense physics can't be described fully by reference to "abstract laws", that doesn't mean that we should just give up. In the real world, theories aren't just right or wrong, but provide a greater or lesser measure of predictive competence—and even a partially correct theory is better than none at all.

This proposal is, like Hayes', a descriptive one: the construction of broad and shallow models of common-sense psychology, along the lines of the three aspects sketched out in this chapter, of anthropomorphism, simulation, and theory. Attempts to build models of small parts of common-sense psychology have not been so successful as to indicate that they are necessarily on the right track. At least to start with, a broad and shallow approach is needed to sketch out common-sense

psychology; it is not yet anywhere near as clearly structured into topics as Hayes presents common-sense physics. Pushing hard on one topic, just like an air bubble under the wallpaper, might just move the problems somewhere else. We need to look at the whole problem and at the relationships between the different topics before we can begin to work out the complete structure of common sense psychology.

But this proposal is methodologically different from Hayes' in that working models of human common-sense psychology are to be constructed. There are several reasons for this: first, the psychologists and philosophers have asked for one, and lament artificial intelligence having "lost much of its ambition in this area" (Samet, 1993). Second, in the absence of a working model, theories become too hard to evaluate, especially for common-sense psychology which depends on comparison with people, and where the interactive nature of a model may have a dramatic influence on how people interpret it.

A third and more fundamental reason is that for the higher orders of mental states, a compatibility between psychologies is essential. A statement like 'I believe that she believes that he believes that...' depends on my recognising someone else's beliefs as like my beliefs. If we are to be able to make statements like this about machines—and the Turing test surely permits this—then there must be a *compatibility* between human and machine competence at recognising beliefs in others. Again, I will return to this theme in chapter 13.

Conclusions

The problems that artificial intelligence is tackling are big ones—big enough to make some think that there are fundamental and possibly irretrievable flaws either in the discipline or even in the whole of science. This is something of an overreaction; certainly our anthropocentricity is a big problem, but not one that is inaccessible to science. There is little evidence that, as Dreyfus and Dreyfus (1988) claim, the problem of common sense cannot be solved even in principle, although the apparent lack of progress in this area is a bit disconcerting. While work is progressing, it is hampered by some rather big methodological problems.

At the core there was a simple problem: we forgot about anthropocentricity and took too much of what we intuitively felt to be right as the truth. Stepping outside our humanity is something that perhaps we can never do in principle, but that doesn't mean that we shouldn't try—not by a regress to the Skinnerian vantage point (with apologies to Dennett, 1987) denying human mentalistic terms completely, but by indirectly looking at the effects of the ultimate unobservable, our anthropocentric point of view.

Although there is no justification for a *definition* of intelligence as the behaviour that we recognise as intelligent, any more than there is for any other definition of intelligence—it may be that intelligence is identified principally by our ability to recognise it. Unfortunately, we don't really know—even informally—what intelligence is, although we do seem to have some idea intuitively, and we usually do recognise it when we see it. To find out more about intelligence, we need to look at how people (and perhaps animals, too) ascribe mental states to other people, to animals, to machines, perhaps even to thermostats, all under a variety of different conditions. Artificial intelligence has a role to play here, too; we can use it to build models of this ascription process—models which can be compared to those of people (Samet, 1993). This proposal sketches out how this might be begun, but there is still a lot to be done.

At the end of the day, we all recognise intelligent behaviour when we see it. When we see people, we see them as minds, not just as bodies. When we see computers, we just don't see minds. The difference between people and computers lies in ourselves as well as in them, and if we are to overcome this fundamental anthropocentric asymmetry, artificial intelligence must join up with psychology at least to the extent of finding when and how we see minds. It must begin to study common-sense psychology.

In the next chapter, I will look at the problem of anthropocentricity in more detail, and show that it can reveal more clues to the real nature of common-sense psychology. I will suggest that the phenomenon of anthropomorphism mentioned before is an essential element of common-sense psychology, because it determines when we see a mind in a system, rather than simply a body. In the next chapter, and in the one which follows it, I will develop a far more detailed model of anthropomorphism and the role it plays in common-sense psychology.

Chapter 7

Cats, bats, and anthropomorphism

Introduction

Paul Gallico wrote in *The Silent Miaow:* "'Anthropomorphism' is a word you will encounter a great deal in this book, and you had best know something of its meaning. It means that people ascribe human qualities to things or to animals because they are so conceited they think the world revolves around them, and that the greatest thing on earth is to be a human being". But there's a twist; Gallico's book is written in the first cat singular: the narrator is a cat musing on anthropomorphism.

This criticism of anthropocentricity is a damning one—and one that needs to be taken seriously in cognitive science. Anthropomorphism as a term is rigged for humans, but is the phenomenon itself really restricted to our own species? To assume so is rather human-chauvinist. Why is it that humans see themselves as being at the top of some evolutionary ladder? Is it an extension of what Darwin (1871) saw as "our natural prejudice, and that arrogance which made our forefathers declare that they were descended from demi-gods"? When we study the way that we attribute mental states to animals, why is it that there is such a close relationship between the status on this imaginary evolutionary ladder and the physical, even genetic, similarity to humans?

To a casual observer it might seem that anthropomorphism is a rather quirky phenomenon of only indirect relevance to common-sense psychology. On the contrary, I think it is right at the heart of the problem. One of the most common qualities that people ascribe to animals and to objects is human mentality. Anthropomorphism is part of what makes us see others as minds rather than as bodies, and as such it is of fundamental importance to common-sense psychology.

Of course, there is a lot more to it than that. Although anthropomorphism is intimately related to common-sense psychology, it isn't immediately obvious how it is related to propositional attitudes. And besides, our anthropocentricity introduces all sorts of methodological problems. Common-sense psychology is special because we, the people who study it, have exactly this common-sense psychology, and our perceptions are indelibly tinted by this. One part of our having this common-sense psychology is that we have a tendency to empathise with and ascribe mental states to other people—and to things—which appear to have minds.

Pervasive phenomena like this natural common-sense psychology are hard to study. We can draw an analogy with gravity: although we can counterbalance the effects of gravity to some extent, we can't (with our current knowledge of physics) actually remove it in practice, so theories of gravity are difficult to falsify. Having common-sense psychology is similar, we can't just switch it off and see what cognitive science looks like without it—the best we can do is to try to counterbalance its effects.

But if problems are opportunities in disguise, this one is no exception. Although anthropomorphism may break some of our more simplistic notions of common-sense psychology, it offers us some possibly significant phenomena that can be studied and incorporated into a new generation of theories. And although we can't apparently escape from anthropocentricity, this very inescapability may hint at the problems underlying how we see others as having minds.

In this chapter I will examine the psychology of anthropomorphism and its effects on philosophy and cognitive science. My claim is that anthropocentricity is a natural if inescapable part of human psychology and accordingly, that some approaches to the study of the mind need to be treated with more caution than has traditionally been the case. The question is: why do we think the world revolves around us, and what are the implications of this anthropocentricity?

Anthropomorphic excursions

When I live with a cat, it begins to interact with me as a member of its social group. It begins to use its gestural 'language' with me in part as it would with a member of its own species in its social group in the wild. Even these gestures were at least comprehensible enough to humans for Darwin (1872) to be able to describe them and relate them to corresponding gestures and mental states for humans. Domestication enhances the development of this kind of interspecies interaction; it can provide both the cat and the human with enough familiarity with each other for a common interpretation of the gestural language to be developed. Particularly after this domestication, the cat sees me in two ways: physically it still sees me in my human form and therefore as fundamentally un-cat-like, but somehow it still manages to accept me as something of a cat, and it uses its cat gestural language to interact with me in new ways. This suggests that there could exist, in principle, both anthropomorphism from our human point of view, and the cat equivalent, we might call it ailuromorphism, cats ascribing cat-like qualities to things or to humans. When I interact with a cat, both kinds of anthropomorphism can happen simultaneously; I will anthropomorphise the cat and the cat will ailuromorphise me. Perhaps this example is a little improbable, though; as Eddy *et al.* (1993) say "to date, only well-documented instances of zoomorphism have been demonstrated in chimpanzees". But in these interspecies social relationships, the participants in some sense lose their species boundaries and become chimerae; the cat becomes a superposition of cat and human to me, and to the cat, similarly, I become a superposition of human and cat. The interaction between us becomes neither purely cat nor purely human (Haraway, 1992).

The literature of cognitive science is quite freely populated by animals of various species trying to interact with humans, but perhaps the most famous instances are Wittgenstein's lion (Wittgenstein, 1953) and Nagel's bat (Nagel, 1974). Wittgenstein (1953) wrote "if a lion could talk, we could not understand him". Wittgenstein's point was that using language is part of a "language-game"—part of an active "form of life" including, for example, customs and institutions—outside which that language cannot be understood. He draws an analogy with a country with a foreign language and traditions: "even given a mastery of the country's language, we do not *understand* the people" (Wittgenstein, 1953, original emphasis).

But somehow we intuitively feel that we should be able to interpret, to some very small degree, what a lion might say to us about some things, even though we don't speak 'Lion'. What about paralinguistic communication in the foreign country analogy? Wittgenstein also wrote: "if I see someone writhing in pain with evident cause I do not think: all the same, his feelings are hidden from me" (Wittgenstein, 1953), but can't we also make this connection with the people of this foreign country, and even with animals? We don't usually see lion pain with the intensity we see human pain, but we can and do recognise it, and we consider it something close enough to our own experience for it to be morally wrong not to prevent or alleviate it.

So here Wittgenstein and I might part company: I think it entirely possible for forms of life to overlap to a small extent, and to this extent we can have something like a common language with cats, even if only a language of the most primitive form. After all, we share a common biological inheritance and a similar physical environment, and these have shaped our forms of life, to some extent side by side. Darwin (1872) said: "in the case of lower animals involuntary bristling of the hair serves, together with certain voluntary movements, to make them appear terrible to their enemies; and as the same involuntary and voluntary movements are performed by animals nearly related to man, we are left to believe that man has retained through inheritance a relic of them". And with the action we have retained an interpretation of the action.

Nagel's bat also presents us with a case which touches on human-animal communication but from a very different point of view. His conclusion is described by Akins (1993) as "the only possible access one could have to the phenomenal experience of another organism is by means of a kind of empathetic projection—by extrapolation from one's own case". It is this very projection that makes us intuitively feel that we could understand Wittgenstein's lion, and it is the *empathetic* nature of the projection that shows how essential the psychology of the observer is to this ascription.

The bottom line is this: when we are asked about the bat's subjective experience we can't answer: we can never truly know what it is like for a bat to be a bat, which is what Nagel is asking, but we automatically switch to the analogical form of the expression 'what is it like'. This is so natural we

often forget to ask ourselves how and why we do it, and we find ourselves asserting, as Dennett does, that these analogical narratives should be accepted "tentatively—pending further discoveries—as accurate accounts of what it is like to be the creature in question" (Dennett, 1991).

Hofstadter and Dennett push the point further with an extensive variety of questions in the 'what is it like' format (Hofstadter & Dennett, 1981). Their point is, quite simply, that the same problem arises in other cases where there is a similar barrier to experience. And the problem doesn't go away even if we apply the idea to the first person, with questions to ourselves like 'what is it like to be me when I'm angry?' and 'what was it like to be me fourteen minutes ago?' The difficulty of these shows that the 'what is it like...' format shows some opacity even to one's own previous mental states. Perhaps the only question Nagel could truly know the answer to is 'what is it like to be me now?'

Before moving on, there is one important distinction to be made. There are two different kinds of anthropomorphism shown in Wittgenstein's and Nagel's experiments: we anthropomorphise the lion but we try to share the bat's experience—it is more like identifying with the lion than seeing it as a human. The first kind of anthropomorphism is projective, in which external animals and systems become chimerae through the superposition of aspects of the observer, and the second is introjective, in which the observer comes to be, in part, a chimera with the observed system. In the case of Wittgenstein's lion, or (more clearly) Tenniel's lion illustration in Lewis Carroll's *Through the Looking-Glass* it is the projective kind of anthropomorphism, but when we ask ourselves what it is like to be Nagel's bat, it is a case of the introjective kind, because it is *we* that change, not the bat. These two aspects are closely bound together. As part of projectively anthropomorphising something, I begin to introject what it feels like to be that something. Identification comes bound up with anthropomorphism; it seems probable that in the projective kind, we are using something like a theory, where in the introjective kind, we are using something like simulation—asking ourselves what it is like to be that something.

At first glance it seems that these two different kinds of common-sense psychology, particularly with the correlation between them, might best be considered different sides of the same coin. This, indeed, is effectively the claim of Davies (1994) with regards to the 'theory of mind' research

reviewed in chapter 3. I am not so sure; although it is possible the case for this is not yet proven. Federn (1952) similarly differentiates between two kinds of identification: "they may involve the entire ego or only one part of it. The mechanism operative in either kind of identification is probably a different one" (Federn, 1952). Without doubt, these two kinds of anthropomorphism are closely related. In one of the very few psychological studies of anthropomorphism Eddy *et al.* (1993) found a strong correlation between them, and it seems more than likely that the principles underlying them are closely connected. For both lions and bats the thought experiments they embody are clouded by the same phenomenon: while the thought experiments remain subject to philosophical analysis, they both open the same trap. We see the animals as other than they are and open, through the narrow overlap in our forms of life, a band of human-animal communication which can trick us about the intended point of the argument.

Anthropomorphism also has important methodological ramifications. Using this projection is seen as bad scientific practice, particularly in biology and ethology, but also in cognitive science. Searle (1992), for instance, comments: "prior to Darwin, it was common to anthropomorphise plant behaviour and say such things as that the plan turns its leaves towards the sun to aid in its survival. The plant 'wants' to survive and flourish and 'to do so' it follows the sun". Searle then tries to "rescue" cognitive science by inverting its explanations from their original teleological forms to functional ones, for example "plants that turn their leaves towards the sun are more likely to survive than plants that do not". Surprisingly, he then goes on to say: "it is easy to understand why we make the mistake of anthropomorphising the brain—after all, the brain is the home of anthropos" (Searle, 1992). He's right, of course, and this is what makes the problem so difficult. Is it possible to remove anthropos from our explanations—from our science—when we can't apparently remove it from ourselves?

So while we may sympathise with Searle's criticism of anthropomorphism in cognitive science, it may well be endemic. Eddy *et al.* (1993) note that it is "almost irresistible". Krementsov and Todes (1991) comment that "the long history of anthropomorphic metaphors, however, may testify to their inevitability". And if anthropomorphism is at the heart of our point of view, banishing it won't help cognitive science. It is perhaps because we are totally unable to step outside our

humanity that it is so tempting to regress to the Skinnerian vantage point (again with apologies to Dennett) and deny the mentalistic terminology that connects the behaviour of others to our own experience. If this is the case—if anthropomorphism is something that we can't remove from our science, but is just something that we need to learn to live with, then perhaps we need ways to "set traps for it" (Caporael, 1986). Then, at least, we might be able to recognise it when it happens, and take this into account when forming scientific judgements. I'll come back to these issues in chapter 14, when I'll discuss some of the methodological ramifications of common-sense psychology.

Anthropomorphism reconsidered

Anthropomorphism as a psychological phenomenon has not been seriously studied: only a few have seen it as worth study in its own right, rather than as a hindrance to proper study (e.g. Caporael, 1986; Eddy *et al.*, 1993; Tamir & Zohar, 1991). Eddy *et al.* (1993) carried out a detailed analysis of the patterns of anthropomorphism with respect to four different cognitive dimensions for thirty different animals, varying from humans to worms. The four dimensions subjects were asked to rate were:

(a) the extent to which the animal experiences the world in the same way as the subject (an instance of introjective anthropomorphism)

(b) the extent to which an animal can distinguish between a mirror image of itself and another animal (an instance of self recognition)

(c) the extent to which an animal could deceive another into going to the wrong place for food, in order to get it for itself (an instance of deception)

(d) the extent to which an animal could distinguish between being kicked or tripped over (an instance of projective anthropomorphism)

The subjects were also asked for a judgement of the similarity of each animal to themselves, and their previous experience with animals was also evaluated. The results are shown in figure 7.1.

These results clearly show that the subjects were more likely to ascribe high ratings on all four cognitive dimensions for animals that they considered were similar to themselves. They also show that the 'similar experience' metric was consistently slightly lower than the other ratings. This seems to confirm the comparative difficulty in ascribing consciousness rather than cognition: "many contemporary thinkers are willing to attribute thought to animals, but not consciousness" (Bechtel, 1992). But even the effects of similarity are complex and subtle. For example, Eddy *et al.* found there are big differences between general cases and specific ones: "perceived similarity toward *particular* animals and *clearly defined mental states* can make an important difference in the degree to which people engage in anthropomorphism" (Eddy *et al.*, 1993, original emphasis).

These effects seem to show a persistent and inevitable tension between the simulation (role taking, introjective anthropomorphism, empathy) view and the theory (heuristic, projective anthropomorphism) view, although more empirical research is needed in this area This is an area which is still subject to very active research by people from many different disciplines interested in common-sense psychology, and the relationship between the two is still far from clear. In both simulation and theory, though, the correlation with similarity is still the dominant effect in Eddy *et al.*'s

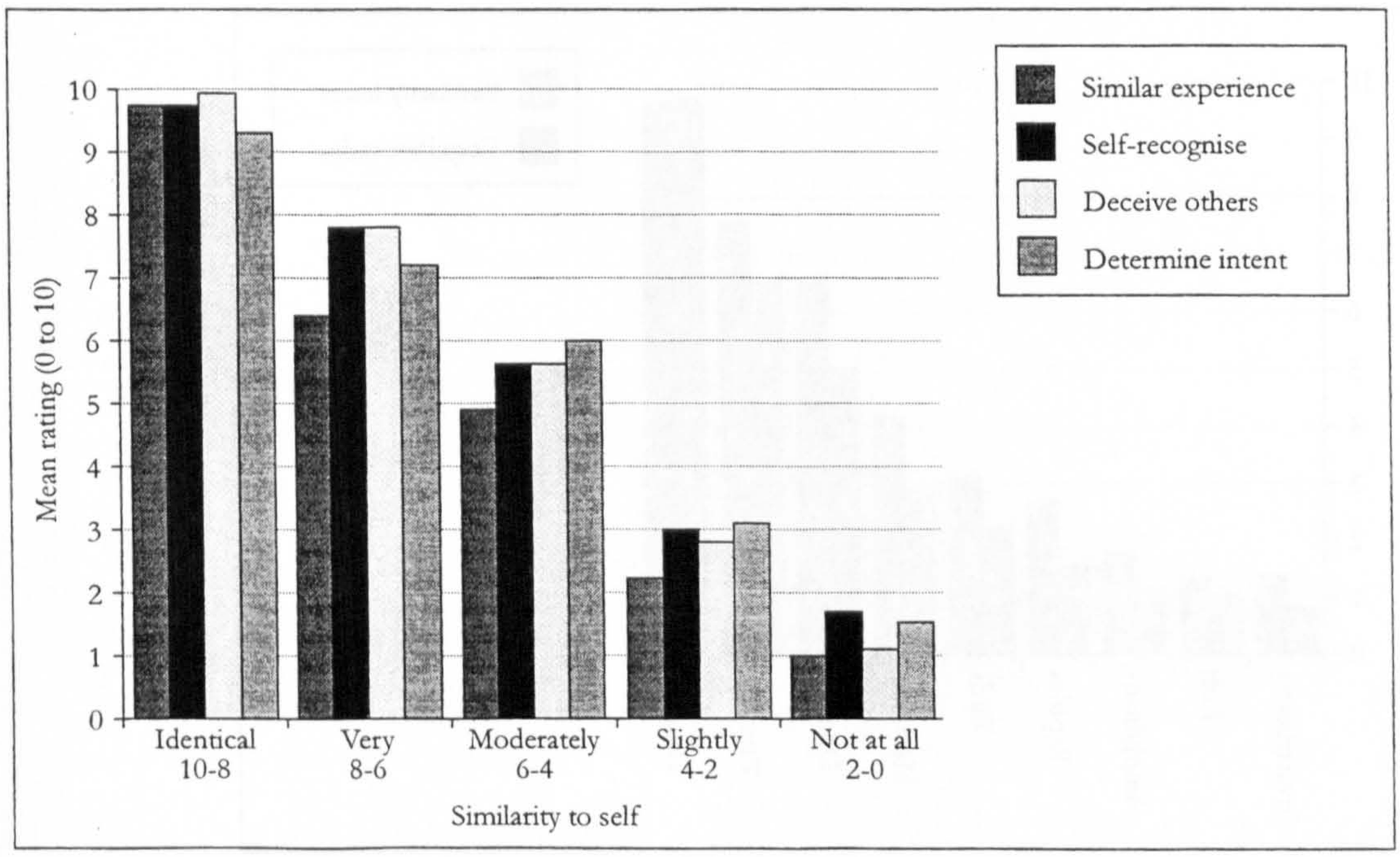

Figure 7.1. Ratings on four cognitive dimensions against perceived similarity (after Eddy *et al.*, 1993)

results. Although there are variations on the theme, then, there is evidence that ascribing both cognitive skills and consciousness are, at least in part, influenced by the principles of common-sense psychology.

The subjects' rating of similarity is only part of the story. The ascriptions can also be compared to the major phylogenetic groups, as an approximation to genetic similarity. Figure 7.2 shows the average rated similarity and the average of the four cognitive dimensions for the phylogenetic groups in the study. Again, the correlation between similarity and the overall cognitive score is good, and there also a strong correlation with genetic similarity, if the relatedness of phylogenetic groups is taken as an approximation to genetic similarity. It is also worth noting the consistent overestimation of cognitive abilities with respect to the estimated similarity.

Eddy *et al.* do relate their study of anthropomorphism to Dennett's (1971; 1987) intentional stance: "Anthropomorphism of cognitive states in our terms will involve attributions of mental processes that are not reflexive in nature (first order intentionality), but are equal to or greater than Dennett's second order of intentionality" (Eddy *et al.*, 1993). But this isn't quite what their experiment tests

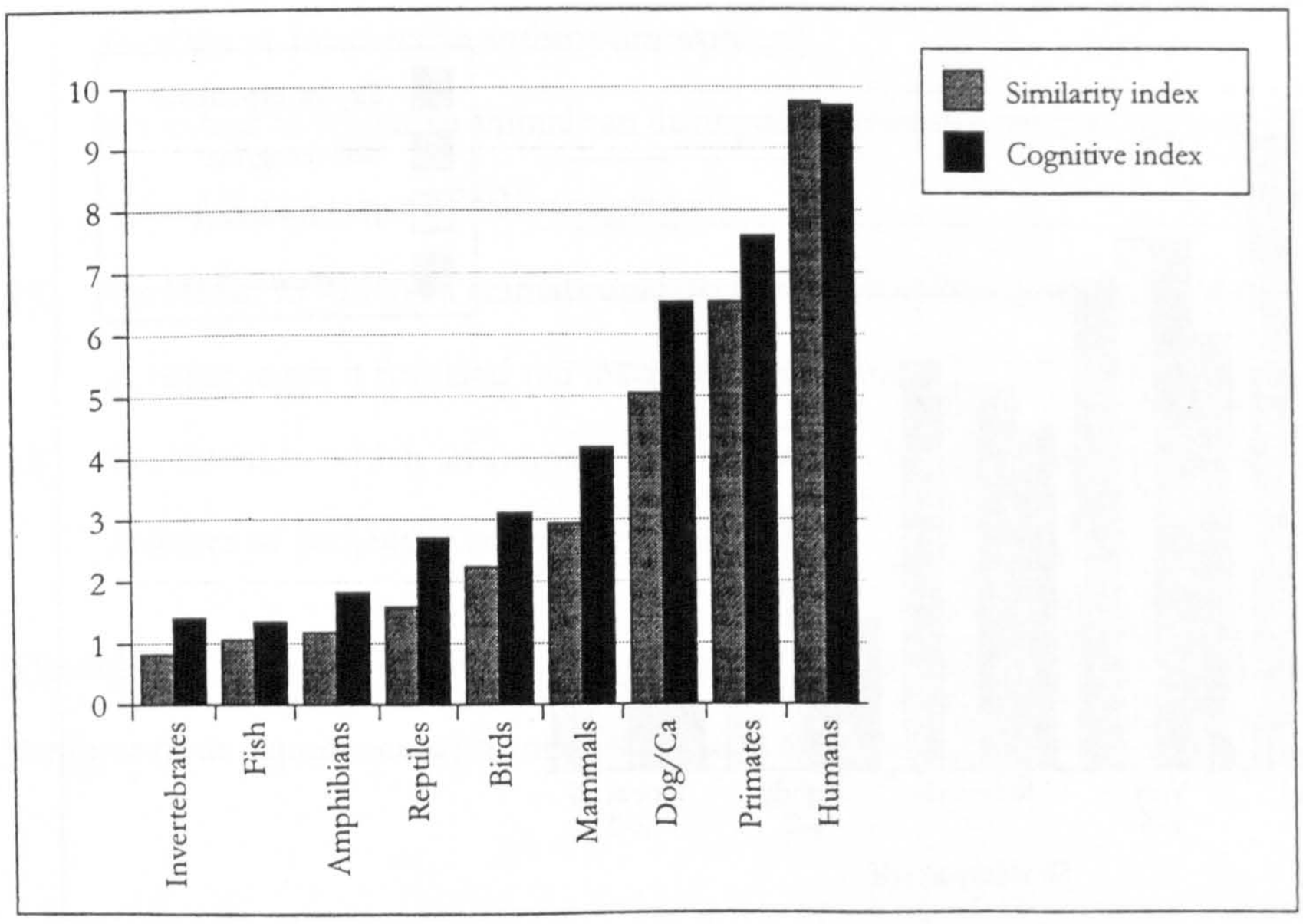

Figure 7.2. Similarity and cognitive scores for the major phylogenetic groups (after Eddy *et al.*, 1993)

for—even the more cognitive dimensions of apparently recognising intentions and deception don't qualify; this was the essence of Dennett's (1978), among others, criticism of the methodology of Premack and Woodruff's experiments (1978), which led to the false belief test.

Some more evidence on anthropomorphism can be gathered from Tamir and Zohar's (1991) rather smaller study of anthropomorphic, teleological, and causal explanations of animal and plant behaviour in education—comparing exactly those kinds of explanation Searle (1992) uses to justify explanatory inversion. Although the subjects for this study were in two groups aged 15 and 17, there was still evidence of a "gradual development of reasoning patterns" indicating that anthropomorphism as phenomenon isn't rigid—although there is no evidence from this study on any of the factors influencing this development. They also found a clear distinction between two kinds of teleological reasoning, one of which was closely correlated with anthropomorphism, while the other is based more on functional reasoning. This seems to back up to Dennett's distinction between the intentional stance and the design stance.

Caporael (1986) provides an extensive and insightful theoretical analysis of anthropomorphism—including a description of how fundamental a bias it is for people. Caporael interprets the phenomenon of anthropomorphism as a "'default schema' applied to non-social objects, one that is abandoned or modified in the face of contradictory evidence". Unfortunately, this just doesn't match Eddy *et al.*'s data, in that the familiarity aspect seems to amplify the phenomenon rather than to reduce it. Besides, her model assumes that animals and computers are 'non-social', and the evidence is against this being true (Haraway, 1992; Nass *et al.*, 1994). Animals and computers probably are genuinely social, but neither to the same extent nor in the same way that people are.

Even so, Caporael's analysis has much to offer. She shows that the anthropomorphic bias fundamentally undermines the Turing test: "if there is an inherent anthropomorphic bias, the imitation principle clearly incorporates a confirmatory bias, and is therefore unreliable" (Caporael, 1986). I've discussed exactly this problem, and described a few examples, in chapter 5's review of the Turing test. Caporael then suggests that if the only ways of denying this bias (accepting one or another form of solipsism or rampant relativism) are ruled out, the only alternative is to accept the anthropomorphic bias, but to "set traps for it" (Caporael, 1986).

The approach I want to take is to interpret anthropomorphism as a phenomenon of a common-sense psychology. That is, common-sense psychology isn't just restricted to the processes we use to reason about other people's mental states, it also contains an aspect which ascribes them the ability to have mental states in the first place. Anthropomorphism, as a tendency to preferentially ascribe mental states to some things over others, should fall within the realm of common-sense psychology, and, therefore, its patterns will reveal patterns in the underlying common-sense psychology.

Armpits and anthropomorphism

As I've already mentioned, there has been very little research on anthropomorphism. This neglect means there are few theories to help explain the origin of anthropomorphism in any detail. The theories which do exist have a number of features in common: they all take anthropomorphism as natural and adaptive, and as probably not restricted to the human species. They also see a connection with altruism: for instance, one possibility indicated by the apparent importance of phylogenetic similarity is Dawkins' (1982; 1989) analysis of genetic altruism: "if an altruistic animal has a cake to give to relatives, there is no reason at all for it to give every relative a slice, the size of the slices being determined by the closeness of the relatedness. Indeed, this would lead to absurdity since all members of the species, not to mention other species, are at least distant relatives who could therefore each claim a carefully measured crumb!" (Dawkins, 1989). Oddly enough, this seems to be exactly what is going on in anthropomorphism, closely matching the results of Eddy *et al.* (1993) in figure 7.2. But for this to be adaptive a gene needs to be able to 'recognise' copies of itself in others; how can this be? One possible mechanism is a "green-beard effect" (Dawkins, 1989; Hamilton, 1964), a genetic linkage with two effects, a label (such as a green beard), and a behavioural pattern which includes acting altruistically to individuals with that label. While this linkage will be favoured by selection, it is vulnerable to outlaws who skip the behavioural altruism but keep the label, so that they still receive the benefits of it from others. Dawkins' view is that the green-beard effect is "far-fetched" (Dawkins, 1982).

Instead, he suggests that an "armpit effect" can masquerade as a green-beard effect, using self-comparison to bypass the need for a label: "in the paradigm hypothetical example, the animal is supposed to smell its own armpits, and behave altruistically towards other individuals with a similar smell" (Dawkins, 1982). The altruistic behaviour isn't triggered by a definite label, an absolute property of an individual, but by one animal's perceived similarity of another, which is, therefore, relative between individuals. Is anthropomorphism, then, an armpit effect? Eddy *et al.*'s evidence shows a gradual trend of increasing anthropomorphism as similarity increases, indicating that an armpit effect is a more likely cause than a green-beard effect. The kind of similarity comparison in anthropomorphism allows continuous variation between individuals, where the presence or absence of a label isn't open to such variation. Eddy *et al.* also allude to this kind of possible genetic basis for anthropomorphism: "it may be that the extent to which we share genes in common with the animal in question increases the likelihood with which we perceive them as being endowed with comparable cognitive abilities" (Eddy *et al.*, 1993).

I should point out, though, that the difference between an armpit effect explanation and a green-beard effect explanation is not as significant as it might seem. The armpit effect, in its paradigm case, also needs a 'label', but that label (the smell of one's conspecific's armpit) is subject to continuous variation and can be used to measure similarity, rather than simply detect it. Because even armpit effects require something which can be used to measure similarity, the features that are actually used for this in anthropomorphism still need to be investigated. Some probable elements, such as facial topology, seem to be apparent already, but there is still a lot of research to be done in this area.

But genetic explanations only go so far. While this works for lions and bats, it doesn't explain why we still anthropomorphise trains, even though they can't share any of our genes—it is here, perhaps, that Caporael's (1986) 'default schema' interpretation of anthropomorphism seems more attractive. Another possibility is that anthropomorphism is part of a social altruism, rather than a genetic altruism. This is where the connection to common-sense psychology becomes important. Common-sense psychology has gifted animals, especially humans, "with remarkable powers of social foresight and understanding" (Humphrey, 1976), social interactions are transactions, we act

through each other, continually playing games of understanding and "guessing thoughts" (Wittgenstein, 1953). Humphrey's theory goes further: he suggests that the transactional nature of interaction is so persistent that even in situations which aren't transactional we still play our psychological games and see things not as they are, but as they should be if they were players in these games. Humans have naturally "explored the transactional possibilities of countless of the things in their environment, and sometimes, Pygmalion-like, the things have come alive" (Humphrey, 1976). This is certainly an accurate description of anthropomorphism, but for Humphrey this anthropomorphism is almost an epiphenomenon, and he doesn't really deal with the question of *why* we so readily set up altruistic behaviour patterns with non-human systems.

Graham (1987) suggests a third possibility, which extends common-sense psychology with a predator-prey psychology. Imagine a human with an extended common-sense psychology which was capable of predicting not only the behaviour of other humans, but also of animals. For a hunter such as an early human, this would confer a considerable advantage, even if it worked only rarely, and even if it had to be supplemented by learning. Similarly, this same faculty could offer an animal—even a human—an advantage when trying to avoid being eaten. This is not a particularly significant point to make here, but my claim is a simple one: a faculty as powerful as common-sense psychology could be of use to an animal in all the four 'F's, and not just for social interaction. So developed, it could naturally be applied to many different kinds of animal, not just those of the same species.

At this stage, though, it is probably premature to select which of these theories is most likely; and besides, they are not really incompatible and a complete explanation of anthropomorphism may well include elements of all of them. But before we can begin to evaluate these theories more fully, and trace their implications for common-sense psychology, we need a more detailed examination of the behavioural regularities of anthropomorphism. This may show more clearly which kind of theory we need, and perhaps even provide a better foundation for this theory.

The regularities of anthropomorphism

The first two regularities are those found by Eddy *et al.* They suggest that there are two primary factors involved in anthropomorphism: "people are likely to attribute similar experiences and cognitive abilities to other animals based on (1) the degree of physical similarity between themselves and the species in question (e.g., primates), and (2) the degree to which they have formed an attachment bond with a particular animal (e.g., dogs and cats)" (Eddy *et al.*, 1993). These two regularities, which I will call 'similarity' and 'familiarity', are the two most important regularities in anthropomorphism.

Similarity is, broadly speaking, a measure of the physical similarity between one system and another. I am using the term 'system' in a loose way here, to cover everything from humans, cats and bats, to railway trains and rocks. Similarity is a metric distance between systems; that is, a measure of the relative difference between systems. This can be contrasted with the common anthropocentric view which suggests that humans are in some absolute sense the highest possible class of system. This follows from interpreting the similarity aspect of anthropomorphism as indicative of an armpit effect rather than a green-beard effect. The importance of similarity shows that form matters: Turing's (1950) "fairly sharp line between the physical and the intellectual capacities" of people is fuzzy in practice. Eddy *et al.*'s second component is familiarity; this too is a metric distance between one system and another, but this time it measures the amount that one system has learnt to predict and work with the form and behaviour of another.

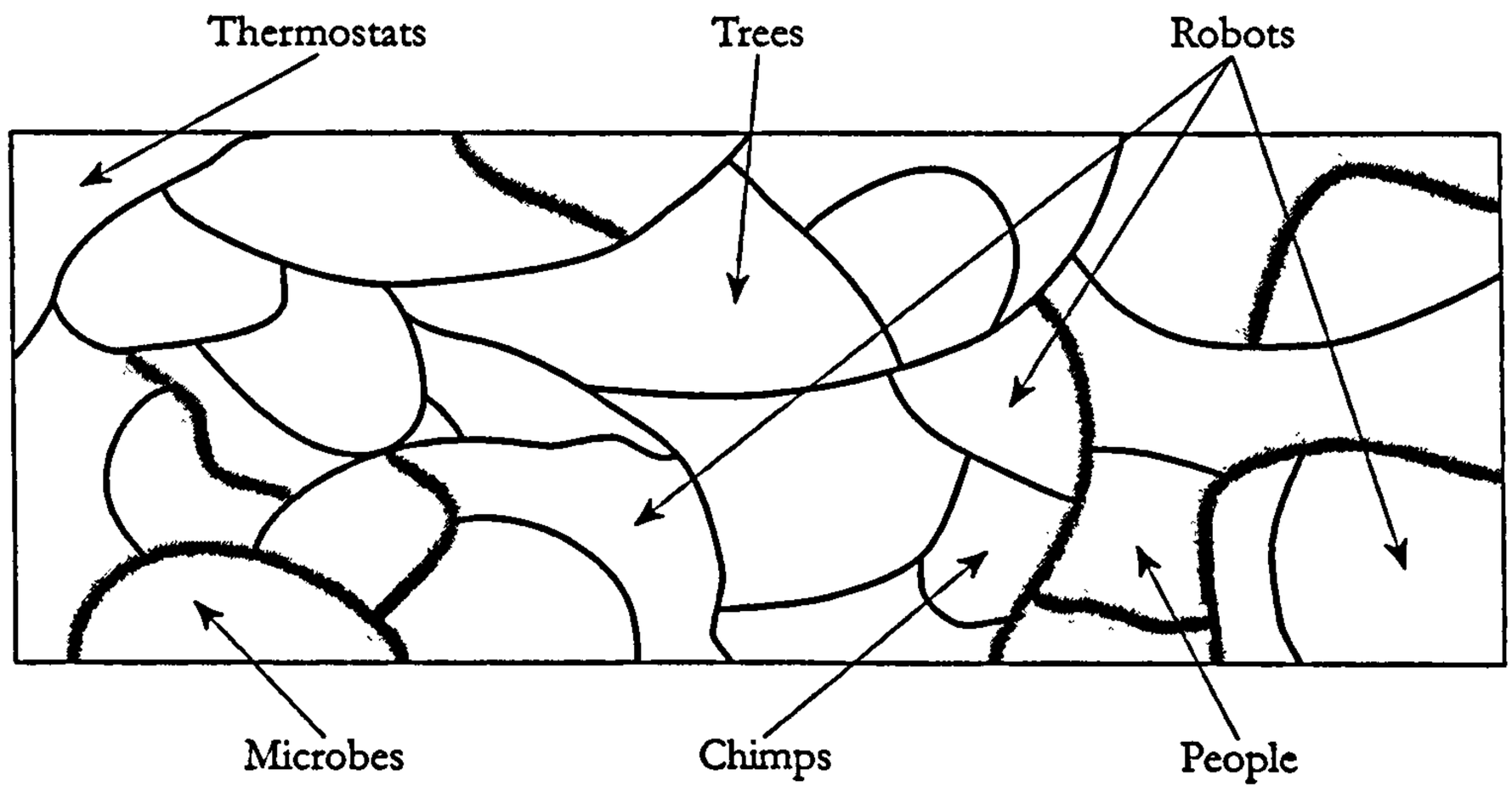

Figure 7.3. The "design space" (after Sloman, 1993)

To show the concept of a metric space, I've borrowed Sloman's (1993) "design space" for figure 7.3. Originally, Sloman used this figure to illustrate that the differences between different kinds of systems can occur in varying degrees of sharpness; some offer continuous variation between systems, where others appear to differ by a sharp discontinuity. There is no absolute distinction between things with minds and without, but nor is there any continuum. The result is a kind of landscape of cliffs and slopes between the different kinds of system.

I want to offer an additional interpretation of this diagram, to reflect the notion of similarity: instead of the landscape offering descriptions of categories, it can be seen as a metric space, where the distances and paths between any two points on the landscape indicate how an instance at one point will be perceived by an instance at another. It is not the absolute position or the absolute properties of a single region that matters, but the relative distance between positions.

Besides similarity and familiarity, there are a whole gamut of subsidiary regularities which overlay these. Animation, for instance, is very important. Although it is possible to anthropomorphise things which don't move, it is almost impossible not to anthropomorphise things which do move, to some extent at least. Trains, aeroplanes, clouds and vacuum cleaners, we see faces in them all, sometimes even going to the extent of painting them on just to make it really obvious. Other subsidiary regularities associated with the form indicate that the social context and richness of the interaction medium have an effect on anthropomorphism; these showing up clearly in studies of the Turing test (Turing, 1950), as I discussed in chapter 5. And the kind of prediction, the reason *why* we are taking a stance to the system, is also an important influence on the stance selected.

Finally, there is one curious but very important regularity: the more I think or know about the mechanism of something, the *less* I am able to anthropomorphise it. As Woolgar (1985) pointed out, whether devices can be called 'intelligent' is at least partly to do with whether we know how they work. We can call this regularity 'structure', showing that the ascriptions depend on a compatibility between our understanding of the underlying form and the mentality that can be ascribed to the system as a whole. Taking anthropomorphism as a tendency to take the intentional stance (Dennett, 1971), it is as if being able to take either the design stance or the physical stance inhibits the natural anthropomorphism; perhaps this shows the affinity with Caporael's (1986)

default schema interpretation. Dennett's stances have psychological aspects as well as philosophical ones, and more importantly these psychological aspects interact with one another, being to some extent mutually exclusive. The structure factor is a strange regularity for anthropomorphism, but an important one with some wide implications which I will discuss more in the next chapter.

To sum up, then, I have identified seven important regularities for anthropomorphism, listed in table 7.1 below. These regularities correspond to the regularities of Eddy *et al.*'s experiment, and the patterns found in other studies of the ascription of mentality, including, for example, the review of the Turing test in chapter 5. Of these, only similarity and familiarity have been looked at in any detail so far. I will look at all the factors in more detail in the next chapter.

Are these regularities enough to build a theory of anthropomorphism? There are clearly a lot of questions that remain unanswered, in particular, whether it is an armpit effect or a green-beard effect. In birds, the apparently related phenomenon of imprinting indicates that animation is of innate importance, and although size, shape, and colour are not so relevant as in human anthropomorphism, they do have an effect (Hinde, 1974). Bowlby (1969) sees related patterns in human attachment processes. Of course, all armpit effects involve some behaviours from birth: a disposition to sniff armpits is needed even for the paradigm case. Armpit effects, leaky common-sense psychology, and a predator-prey psychology all show some of the characteristics of anthropomorphism, while none is sufficient to explain it completely. But the theories aren't really incompatible, and a synthesis of the three may well prove a good initial basis for more complete models of anthropomorphism.

	Factor name
(a)	Similarity
(b)	Familiarity
(c)	Animation
(d)	Structure
(e)	Interaction medium
(f)	Social context
(g)	Predicate

Table 7.1. Factors involved in the ascription of mentality

But modelling anthropomorphism in common-sense psychology is only part of the story. A better understanding of anthropomorphism will add to our understanding of human psychology, but its effects run deeper than that. Anthropomorphism is involved in our day to day judgement of other people's mental states, and this is also central to the philosophical and psychological analyses trying to gather together the threads of the study of mentality. In the next section, then, we will look at some of the effects of anthropomorphism on philosophical arguments and psychological mechanisms.

The rationality assumption revisited

One of the criticisms made against Dennett's views in chapter 2 was his strong dependence on an assumption of rationality: "in short, we treat each other as if we were rational agents, and this myth, for surely we are not all that rational—works very well because we are *pretty* rational" (Dennett, 1987, original emphasis). Dennett actually distinguishes two alternative interpretations of intentional vocabulary, a "Normative Principle, according to which one should attribute to a creature the propositional attitudes it 'ought to have' given its circumstances, and one or another Projective Principle, according to which one should attribute to a creature the propositional attitudes one supposed one would have oneself in those circumstances" (Dennett, 1987).

It should be clear that Dennett's Normative Principle is really a philosophical manifestation of the theory theory, and similarly, his Projective Principle is a manifestation of the simulation theory. This duality of philosophical and psychological positions will play an important role in our understanding of the rationality assumption. The tension between the theory theory and the simulation theory raises its ugly head once more. Unfortunately, there are few clues in the dialogue about whether a Normative or Projective interpretation is to be preferred with regards to rationality. Dennett himself avoids taking sides, although he happily categorises others. In fact, Dennett claims that the difference between the two is at most one of emphasis (Dennett, 1987), in that whether a simulation or a theory is to be generated, much the same kind of structure is needed. In this claim, he is not alone (Davies, 1994; Heal, 1994).

As I've already mentioned in chapter 2, Dennett's view of rationality is both evolutionary and pragmatic: it is simply that those systems which weren't rational would have been eliminated by natural selection, so in our day to day interaction we can assume that we won't come across any 'rogue' systems without rationality. And, as I've already mentioned, this doesn't deal with artifacts which haven't been subject to natural selection. In fact, Dennett declines to define rationality at all; he far prefers to stick to a "pre-theoretical concept of rationality" which "relies on our shared intuitions" (Dennett, 1987), and it is this which opens an apparent link to anthropomorphism. Of course, anthropomorphism isn't the same thing as rationality, and I don't want to claim that it is. That doesn't matter, the question is: can it—or does it—play the role of the rationality assumption in Dennett's model.

It seems plausible that anthropomorphism, as a psychological effect, is playing exactly the role of Dennett's rationality assumption. This might, of course, undermine Dennett's philosophical position—not that this is necessarily a problem for a psychological theory derived from it. His claim that the rationality assumption is "pre-theoretical" (Dennett, 1987) allows us to appropriate it as a psychological notion instead of a philosophical one. Can this view be defended? I think so. It is important to remember that Dennett's theory can be taken on two levels. Throughout most of this thesis, and in common with many psychologists, I have taken Dennett's intentional stance to be more or less synonymous with a faculty for common-sense psychology. In practice, the intentional stance can also be taken as a 'philosophicalisation' of this; that is, Dennett's stances might actually be common-sense faculties turned into rather naturalistic philosophical ways of looking at a system.

Dennett doesn't say very much about anthropomorphism, but what he does again appears to draw a connection between it and the intentional stance. In this sense, Dennett's rationality assumption may be, once again, a philosophical manifestation of anthropomorphism, which plays the actual psychological role of a disposition to take the intentional stance. As a parallel, then, this seems to be justified. It is also worth noting that this example hints at predator-prey psychology as a possible influence on the development of anthropomorphism (Graham, 1987).

"And yet this anthropomorphising way of organising and simplifying our expectations about the frog's next moves is compelling and useful. Treating frogs, birds, monkeys, dolphins, lobsters, honeybees—and not just men, women, and children—from the intentional stance not only comes naturally, but also works extremely well within its narrow range. Try catching frogs without it" (Dennett, 1987).

The distinction between the theory theory and the simulation theory in anthropomorphism is really quite a subtle and complex one. As far as the experiments are concerned, ascription of subjective states—which are the cases where the simulation theory usually wins—shows very similar patterns to the ascription of cognitive states, although usually to a slightly lesser degree (Eddy *et al.*, 1993). On this basis, I will interpret anthropomorphism as a disposition to take the intentional stance, and I shall use it as a replacement for the rationality assumption in constructing models of taking the intentional stance. I will discuss the issues involved in this in rather more detail in the next chapter, which looks more carefully at the relationship between the rationality assumption and Dennett's stances.

How does this interpretation measure up against the psychological data presented earlier? Well, if we specify that the strength of the disposition is dependent on our similarity, familiarity, and the various other factors then it does indeed seem to match the psychological data in Eddy *et al.*'s and Tamir and Zohar's experiments. This interpretation is, of course, very different from Caporael's (1986) proposed default schema, but it seems to account for the pervasiveness of anthropomorphism in a way that the schema notion doesn't.

Arguments and mechanisms

Armpit effects work by a kind of analogy, and the similarity component of anthropomorphism clearly shows this character. I have already argued, in chapter 3, that the importance of similarity hints at a deep connection between the psychology of analogy and common-sense psychology. It is as if the 'Argument from Analogy' approach to the 'other minds' problem has been resurrected, but as a mechanism rather than an argument. It is not an argument, it is simply a faculty that people have. Anthropomorphism is how people know that others have minds in practice, in the

real world. Common-sense psychology and anthropomorphism enable humans to go around ascribing minds to things that look and behave as if they have minds. The fact that the argument from analogy is philosophically weak (Hauser, 1993) won't stop people actually doing it.

This tension between the behaviour of a scientist's—or even a philosopher's—common-sense psychology and the structure of arguments, models, and thought experiments is an important theme in this thesis. For example, many thought experiments are liable to a kind of misdirection of intuition: when we study them we need to be aware that they may be triggering our own common-sense psychology so we see the argument in our own psychological terms rather than its own logical ones. Later, in chapters 11 and 12, I will show the extent of this redirection in the model of intuition in Searle's (1980) Chinese Room thought experiment. Arguments have the same problem, as Humphrey suggests: people "expect to argue with problems rather than being limited to arguing about them" (Humphrey, 1976). Academic study of anthropomorphism and common-sense psychology offers us a way of finding the way between the logic of an argument and the intuitions of our own common-sense psychology. If we know how and why we are disposed to see things as other than they are, we can ensure that our psychological models and theories aren't biased by this. Even if we can't step outside our humanity, we can at least begin to become aware of our anthropocentricity and its probable effects on our point of view.

This conception of anthropomorphism is crucial: it tells us that the behaviour that we call 'intelligent' is largely observer-relative. It depends in great measure on the psychology of the observer as well as the actual behaviour of the system. As the regularities suggest, some behaviours and other features of the system will assist ascription, while others will resist it. More detailed study of what promotes ascription and what doesn't is essential to a full study of the mind. It is this recognition of intelligence that is so bound up with anthropomorphism in the psychology of the observer. Our study of psychology needs to be structured around the behaviours that we recognise, and how we recognise them, as much as it is around the behaviours themselves. They are matching lock and key, and neither alone is sufficient, for it is the one that unlocks the other.

Conclusions

To date, none of the disciplines in cognitive science have taken anthropomorphism as seriously as they should. I think this is an error. In part, anthropomorphism is important because it raises a number of methodological issues relating to arguments, models, and thought experiments; I will return to these topics in chapter 14. But anthropomorphism also has more direct theoretical implications which challenge fundamental assumptions in some parts of cognitive science.

Perhaps the most serious of these springs from anthropomorphism's introduction of relativism into the ascription of mental states. If our tendency to ascribe mental states depends on something as superficial as similarity of appearance—and not only on the actual behaviour of a system—then the whole idea of artificial intelligence as something separate from human intelligence is undermined. Even though McCarthy may not care about whether artificial intelligence is psychologically valid or not, systems which pretend to intelligence in a human-like way may be necessary, simply because if they showed an alien kind of intelligence, we would find it much harder to see it as intelligence. This attacks the whole notion of "intelligence in general" (French, 1990). These constraints on our ability to see intelligence are an important theme in the fourth part of this thesis, and will be discussed in much more detail in chapter 13.

It is perhaps curious how little anthropomorphism has come in for scientific study, considering how far-reaching its consequences appear to be. When we look at thought experiments in the literature some throw a light of sufficient contrast that it seems the gulf between humans and animals is unbridgeable, but in reality these experiments often appeal too strongly to our intuitions to be on safe ground. When these intuitions are studied in more detail, it begins to look as if humans have a strong psychological tendency to favour reciprocal altruistic relationships with other humans and animals to a set of clear patterns. Surprisingly, one of the most dominant factors which influences this is perceived phylogenetic similarity. Investigation of possible evolutionary bases for this remarkable pattern of ascription to human and non-human systems shows a number of possible theories for the origins of this kind of reciprocal altruism. Although none

alone appears sufficient to explain all the patterns of anthropomorphism, all are evolutionarily adaptive and plausible; a synthesis of these different mechanisms is viable, and would probably show most of the regularities that we would expect of anthropomorphism.

This study of anthropomorphism reveals a deep issue in cognitive science: there seems to be a bias in favour of seeing others as behavers, to the exclusion of seeing them as recognisers of behaviour. Behaviour and the understanding of that behaviour are intertwined in the continual interchange of actions and perceptions between people in a social context. By emphasising the behaviour over the recognition of that behaviour, we have created an asymmetry in the discipline and are missing out on the study of some complex and important human psychological phenomena. An awareness of the gaps in our minds can help us focus on which behaviours are important in eliciting the right responses from others. This might not seem like much of a goal now, but analyses of the processes underlying why we see others as minds and not just as bodies could prove to be of real significance to the long term future of cognitive science.

But before we can take these ideas further, we need a more precise model of the different factors that go to make up anthropomorphism, and which influence when we take the intentional stance. In the next chapter I will refine this theory of anthropomorphism to the point where it will be possible to build a computational model of anthropomorphism in common-sense psychology.

Chapter 8
Taking a stance

Introduction: common-sense psychology and Dennett's stances

In the previous two chapters, I alluded to Dennett's (1971) notion of the intentional stance in the context of common-sense psychology, and suggested that Dennett's distinction between the intentional stance and the physical stance was rather similar to the difference between common-sense psychology and common-sense physics. In this chapter I'll develop idea this further, looking at the different stances and the different dispositional factors in more detail—enough detail to form the core of the models to be developed in the next part of the thesis.

Dennett's philosophy is based on a distinction between the different stances that we can take towards a system. Dennett (1971) outlines three main stances we can take when, as scientists, we want to understand a system: the "physical stance" which interprets the system in terms of structural objects and physical relationships, the "design stance" which interprets the system in terms of functional objects and relationships, and the "intentional stance" which interprets the system in terms of intentional objects and relationships. According to Dennett, attributing mental states—and, therefore, a mind—is an aspect of taking the intentional stance.

These three stances are not proprietary to philosophy, they describe some of the different ways that scientists can think about systems; but more than that, stances like these are part of our common-sense reasoning about systems. When we try to understand the behaviour of a system, we generally make the interpretation from one of these different stances. The stance that is chosen may be influenced by the interpretation that we are trying to make, but it will not necessarily be determined by it. It is here that anthropomorphism steps in; among other things, it influences which stance is chosen.

In this chapter I will try to defend two hypotheses. The first is that Dennett's stances are manifestations of the different kinds of common-sense reasoning, and the second is that anthropomorphism plays the role of the rationality assumption in practice. In this sense, anthropomorphism is a disposition to take the intentional stance. This possibility has certainly been alluded to by those researchers who have looked at anthropomorphism in detail, as discussed in the previous chapter. The key claims here are that firstly, anthropomorphism is dispositional, and secondly, that the greater the degree of anthropomorphism the more probable intentional reasoning will be.

These hypotheses move parts of Dennett's model of the three main stances from the philosophical realm to the psychological one, as shown in figure 8.1. The main reason Dennett's philosophical position has been selected for this is that, like common-sense psychology, it is explicitly centred in an individual person's ascription—it is observer-relative and does not attempt to discuss 'intrinsic' psychological properties, as do Searle's and Fodor's positions discussed in chapter 3, for example. This avoids the worst problems of anthropocentricity, and is a better match to the model of anthropomorphism developed in the previous chapter.

In psychology at least, some of the patterns of correspondence in figure 8.1 seem to have a wide, if implicit, acceptance (e.g. Baron-Cohen *et al.*, 1985; Wellman, 1990). As psychologists, they, and others, accept a correspondence between the intentional stance and common-sense psychology.

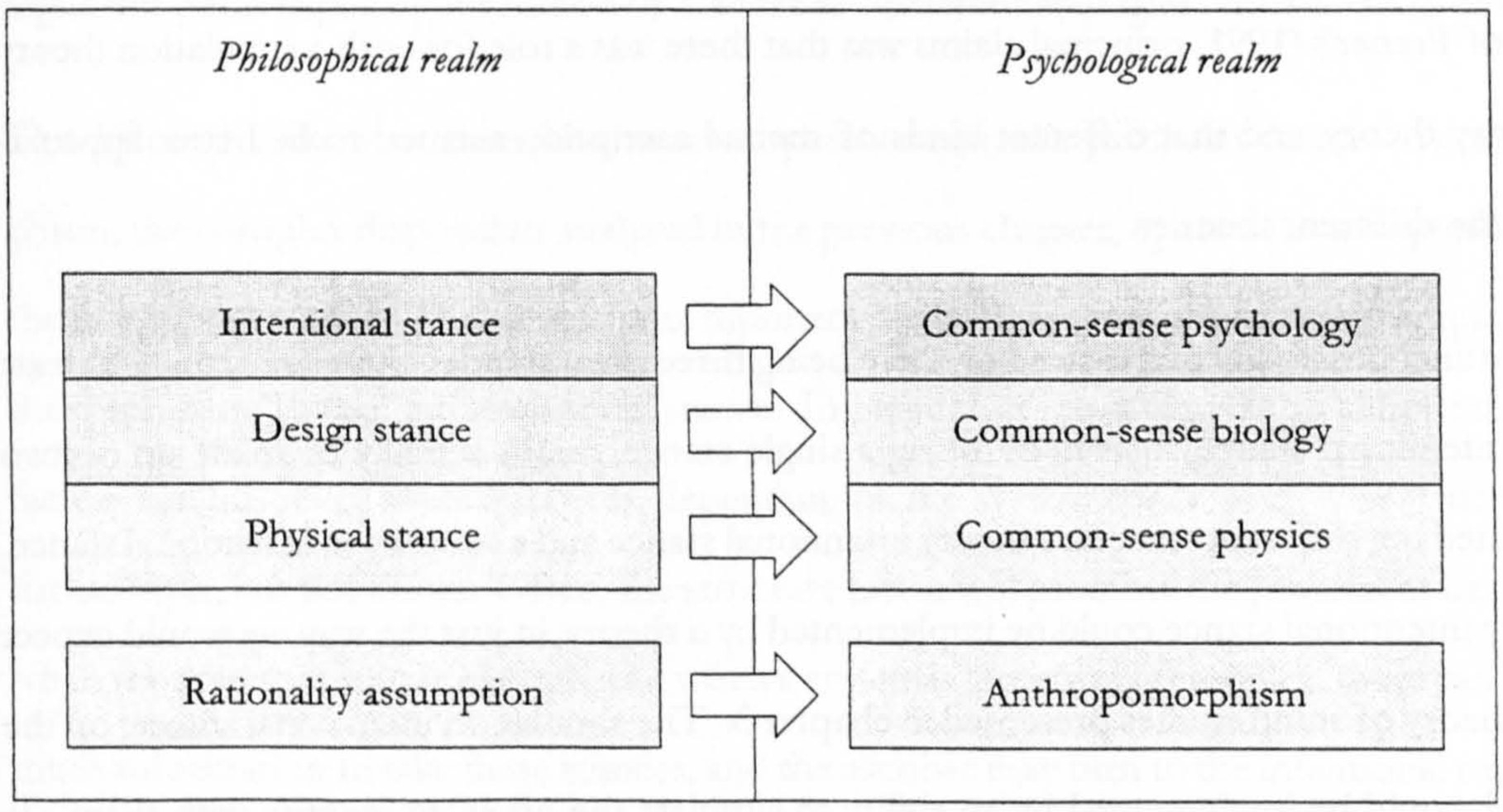

Figure 8.1. Correspondence between the philosophical and psychological realms for Dennett's model

The other correspondences are perhaps more open to debate. I have already shown some evidence for the correspondence between the rationality assumption and anthropomorphism in the previous chapter. The other correspondences—between the physical stance and common-sense physics and between the design stance and common-sense biology—are perhaps the hardest to defend. There is evidence that children do have an innate ability to construct a common-sense physics (Carey, 1985)—although Carey also argues that, in fact, common-sense biology emerges from common-sense psychology, rather than being an innate competence in its own right.

In this thesis, I have effectively left out common-sense biology—although there are occasional elements that do seem to provide some insight into its actual role. The kind of causal and functional reasoning, for example, by which people class things as 'animate' I have left vague. This area has already been studied—and indeed modelled—in some detail by Shultz (1991), but Shultz's analysis corroborates Carey's (1985) suggestion that common-sense biology develops later. In short, the jury is still out on whether there really is such a 'faculty' as common-sense biology, and whether it is distinct from its psychological and physical counterparts.

Are there really four stances?

But is it right to take the intentional stance as monolithic? Some of the evidence from the psychology of common-sense psychology presented in chapter 3 indicates that this might not be the case; one of Perner's (1991) principal claims was that there was a role for both a simulation theory and a theory theory, and that different kinds of mental ascription seemed to be better approximated by the different theories.

So it is a distinct possibility that instead of there being three main stances, there are actually at least four; the intentional stance, instead of being a single stance, might actually be made up of two closely related but different stances: a theory intentional stance and a simulation intentional stance. The theory intentional stance could be implemented by a theory, in just the way we would expect from the theory of mind studies presented in chapter 3. The simulation intentional stance, on the other hand, could be implemented by an ability to simulate the other, to identify with them. Is there any evidence for this claim that the intentional stance isn't monolithic?

As evidence for these two different kinds of intentional stance, there is Perner's (1994) analysis, showing that some behaviours are better represented by a theory theory, and others by a simulation theory. On the other hand, on the philosophical side, there is Dennett's distinction between normative and projective intentional language, which, he argues, "is at most a matter of emphasis" (Dennett, 1987). But as I discussed in chapter 3, sorting out the relationship between theory and simulation is an area where current research is particularly active and particularly controversial, so for the sake of this thesis I will make the simplifying assumption that there is a single, monolithic intentional stance. When I come to model common-sense psychology, though, I will explore the suggestion made by Dennett (1987) and Davies (1994), among others, that the difference between theory and simulation may be illusory.

Modelling the stance taker

The stance taker is the first part of our model for common-sense psychology. It is represented as a competitive system which is subject to a number of influences. Each of these influences acts as a tendency to switch on one particular node in the network, where each node corresponds to a different stance. When the node is switched on, the stance is taken, and the reasoning apparatus which it provides is brought to bear. This representation does depend on the simplifying assumption that the different stances are mutually exclusive. Although Dennett (1971; 1987) doesn't argue this point explicitly, his notion of stances as "different strategies" does seem to bear this out.

The different influences, therefore, act as dispositions to take particular stances. Anthropomorphism, the complex disposition analysed in the previous chapter, operates as a disposition to take the intentional stance. The similarity component of anthropomorphism, for example, increases the disposition to take the intentional stance. In practice, it is not always this simple: most of the factors can influence several stances, depending on the system concerned. The structure factor, for example, has this effect. Often, the structure factor will promote the physical or design stance, when the system is simple enough, but when a system is structurally complex, there can be just too much information to take these stances, and the ascriber may turn to the intentional stance. "The

best chess-playing computers these days are practically inaccessible to prediction from either the design stance or the physical stance; they have become too complex for even their own designers to view from the design stance" (Dennett, 1971).

In developing this model, I will separate the actual dispositions and influences on the various individual stances from the mechanism that selects between the different stances; that is, I will effectively separate anthropomorphism—as a disposition to use common-sense psychology—from common-sense psychology itself. The dispositions corresponding to the various factors generally belong to the individual stances, but before we can be more specific about their relationships with the different stances, we need to study them in a bit more detail.

Looking at the factors in more detail

In this model, most of the factors involved in anthropomorphism, such as similarity, familiarity, and animation, are linked to a tendency to take the intentional stance: this follows from my description of anthropomorphism as a disposition to take the intentional stance in the previous chapter. Some of the factors, though, and notably the structure factor, have effects that are considerably more subtle than a simple tendency toward intentional explanation. In this section I will look in a bit more detail at each of the different factors influencing which stance is selected.

Similarity

Previously, in chapters 5, 6, and 7, I have shown the fundamental role of physical similarity in the ascription of mentality. This is especially clear in the phenomenon of anthropomorphism through Eddy *et al.*'s experimental results. In its simplest form, the similarity effect just behaves as if people tend to take the intentional stance to animals and objects to the degree to which they physically resemble themselves. This interpretation of the similarity disposition is based on the armpit effect (Dawkins, 1989) model of anthropomorphism discussed in the previous chapter.

For the purpose of this model, the similarity measures will be developed using principles derived from numerical taxonomy (Sokal & Sneath, 1963). Numerical taxonomy is a field which has developed statistical measures of the similarities and differences between elements, and techniques for combining these measures into graphs of the relatedness between these different elements. Using a statistical measure like this has its problems; for one, there is no causal element to the model, but since this is intended to be a descriptive rather than a causal account, that is not a problem. A more serious criticism of this approach is that it flies in the face of biological, psychological, and evolutionary plausibility—and there is some ground to this criticism. It really is pretty unlikely that inside a person's head anything like these statistical measures are actually in use. There are other approaches to estimating similarity (e.g., neural network models) but these require training; they are closer to Eddy *et al.*'s familiarity component than similarity. Estimating similarity seems to be a fundamental and widespread psychological effect, not only in anthropomorphism, but, for example, in categorisation, and in the psychology of metaphor, as discussed in chapter 3; so it is quite likely to have an innate side.

There are three steps to using numerical taxonomy (Boyce, 1969; Sokal & Sneath, 1963). The first is to map the different elements to be compared into a multidimensional 'character space', where the axes correspond to all the features which can be measured. This done, the next step is to define a metric on this space—a function that, given any two points in this character space, provides a measure of the distance between them; the smaller the distance between any two points, the greater the similarity between the corresponding elements. Finally, the last step is to combine the points with their closest neighbours in clusters and derive the major groups involved from a substantial sample of points within the character space. For this model, we can dispense with this third stage and rely on the first two—building a metric space which maps elements to points in a multidimensional space and provides a measure of the distance between any two points.

Figure 8.2 shows an example character space, for forms with two features, one on each dimension. Many different distance metrics can be used, and each has different properties. There is the simple Euclidean distance metric, for example, $\sqrt{(X_{j1}-X_{j2})^2+(X_{i1}-X_{i2})^2}$. There is also the

Manhattan distance metric, $|X_{j1}-X_{j2}|+|X_{i1}-X_{i2}|$. Both of these are invariant to translation (additive) changes in the character values. Alternatively, there is the angular metric, θ, which is invariant to scale (multiplicative) changes in the character values.

To set this back in the context of a disposition based on similarity—we can use a character space like this, and its metric, to provide estimates of the similarity between one animal and another, estimates which can then be used to model a disposition to anthropomorphism based on similarity. One point to note, though, is that while it might not seem so at first, this approach is actually anthropocentric, in that the features which are represented are those which are visible to us humans. This is not a serious problem in that these forms are intended to correspond to those found by a perceptual system that will inevitably be human.

In the model, the similarity estimate is actually based on a combination of two different similarity estimates, one for real-valued features, and one for simple attributes which are either present or absent. All of a form's features are divided into these two categories and similarity estimates are evaluated separately for each category, then the estimates are combined and normalised to provide an overall similarity estimate of between zero (dissimilar) and one (identical).

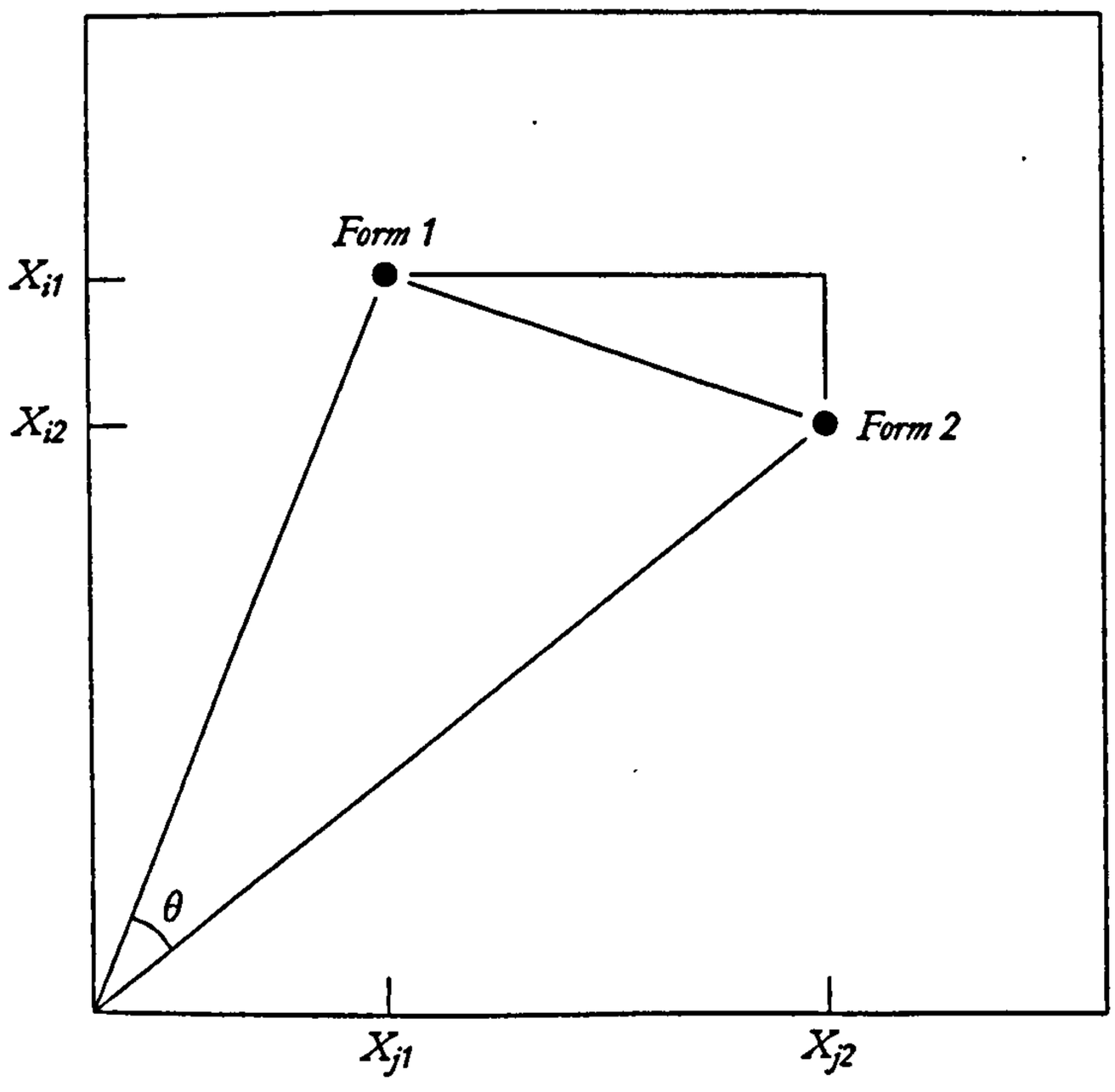

Figure 8.2. An example character space

For the real-valued features, the similarity function used is the product-moment correlation coefficient, which normally has a value between +1 for a strong positive correlation and −1 for a strong negative correlation. (According to the principles of numerical taxonomy, negative correlations are generally improbable, and are a sign of poor data, perhaps using too small a number of characteristics, as one form cannot biologically be the 'opposite' of another; Sokal & Sneath, 1963). The product-moment correlation coefficient has been selected because its classifications seem most like those of human naturalists (Boyce, 1969); and there is a close correlation between the classifications of scientific taxonomy and folk taxonomy (Boster, Berlin, & O'Neill, 1986). The formula for the product-moment correlation coefficient is:

$$r_{jk} = \frac{\sum_{i=1}^{n} (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)}{\sqrt{\sum_{i=1}^{n} (X_{ij} - \bar{X}_j)^2 \sum_{i=1}^{n} (X_{ik} - \bar{X}_k)^2}}$$

where n is the total number of features, X_{ij} is the attribute value for feature i in form j, and $\bar{X}_j$ is the mean of all feature values for form j.

One of the main reasons why this correlation coefficient works so well is because it is invariant with regards to scaling and additive differences between forms (Boyce, 1969), and unfortunately, this is only realistic within a fairly small degree of size difference. Within numerical taxonomy, this is not a problem, because the size differences between closely related animals are generally pretty small, and numerical taxonomy is only used on fairly closely related forms. For a similarity measure in this context, though, it is a potentially significant failing, because it means that an exact miniature of a person would be rated as identical to a real person, at least according to this similarity metric. This is an area where the psychological validity of this part of the similarity metric is questionable. Further research is needed on this, to determine more precisely the structure of people's intuitions regarding estimates of similarity. It is also worth pointing out that similarity estimates are unlikely to be constant for an individual—Carey (1985) shows that there are typically quite dramatic differences between a child's and an adult's similarity estimates for the same forms.

For simple binary features, the similarity function used is the Rogers-Tanimoto metric (Sokal & Sneath, 1963), which is a simple measure of the proportion of features two different forms have in common. This is always a value between zero (dissimilar) and one (identical). The formula for the Rogers-Tanimoto metric is:

$$rt_{jk} = \frac{n-u}{n+u}$$

where n is the total number of features, and u is the number of features that are different between forms j and k.

The values of these two metrics are combined into a single measure of the similarity between two different forms, normalised to a value between zero and one. This creates a metric space on forms which, unlike Sloman's design space in figure 7.3, is basically continuous throughout. Finally, the model actually uses a polynomial modified version of this similarity value: the reason for this is that people seem to be more sensitive to small differences in forms which are close to human than they are to small differences in forms which are wildly different from humans. This is apparent in the non-linear pattern to Eddy *et al.*'s results in figure 7.2. The formula by which the two components of the similarity measure are combined is:

$$sim(i,j) = \left(\frac{rt_{ij} + \frac{1+r_{ij}}{2}}{2} \right)^2$$

Familiarity

The second principal factor is familiarity. Familiarity, of course, implies some kind of learning. This learning, in practice, needs to be unsupervised rather than supervised, ruling out many possible classical artificial intelligence learning techniques, neural networks, and rule induction algorithms. Furthermore, this learning does not appear to be a categorical type of learning, in that the result is dispositional rather than categorical. That is: familiarity affects the tendency to anthropomorphism continuously rather than in discrete chunks.

Because familiarity is learned, there are temporal effects in the learning process that also need to be taken into account. In practice, these seem to differ substantially from one species to another. In most birds, for example, imprinting seems to be closely related to this learned familiarity, yet this takes only a very short learning period and has a strong primacy effect. With human familiarity in anthropomorphism, it is as yet unknown how the familiarity effect behaves in time. Also, familiarity also shows aspects of analogical reasoning; it can be transferred from one individual to other, similar, individuals. When I learn to ascribe mentality to one particular cat, this positively affects my tendency to ascribe mentality not only to that cat, but to other cats, and to other animals which are broadly similar to them, such as lions.

This implies that the familiarity component is, basically, not categorical; that is, it does not depend for its effect on learning categories. This is slightly backed up by the research on animal attachment, which shows similar properties of increasingly ascriptive behaviour, but, in birds for instance, this seems to happen so quickly as to be almost 'one shot' learning which can not provide the generalisation needed to bring in categories. This can't be pressed too strongly as a claim, though, because the effect of familiarity seems to be substantially weaker than the effect of similarity, at least over experiments of short duration. It is also hard to study familiarity because we can't remove it as an effect— among others reasons, because it is unethical. Just as with similarity, our anthropocentricity causes us significant methodological problems with familiarity. This has the effect of clouding its behaviour, and therefore makes it hard to make substantive hypotheses about it without a real danger of getting feet in one's mouth. Even though familiarity does seem to play an important role in anthropomorphism, a lot more study is needed before a more accurate model is possible.

From these results, and from the discussion of familiarity in the previous chapter, we can draw out a number of constraints on the familiarity factor. It should learn, it should learn in an unsupervised way, it should have both a primacy and a recency effect, and it should return a familiarity value on the continuum between 0 and +1. These constraints rule out many traditional learning techniques, such as neural networks (most of which require supervision, most of those which don't are poor at representing real values) and symbolic rule or concept learning systems. Instead,

for the sake of this model, I've adopted a simple evaluation algorithm, which is, basically, a weighted sum of ratings of the quality of ascription, with exponential and decay functions to implement the primacy and recency effects—this is necessary to show the effects of familiarity over a short learning period, such as that afforded by a Turing test. The changes in these primacy and recency effects are shown in figure 8.3. The familiarity at any given time t is:

$$fam(a, i) = \sum_{t=1}^{\infty} r(a, i, t)\, k^t\, e^{-gk^t}$$

where $r(a, i, t) = w\, m\, p_{ait} + (1 - w)\, sim(f_a, f_i)\, p_{ait}$

Here, the familiarity ascribed by a to i is controlled by a rating function, r. There are two components to r; if a has successfully taken the intentional stance to i before, then m will be 1, otherwise 0. This adds an individual rating to r, using the rating p_{ait} as a stored measure of the success of a taking this stance to i at time t. The rest of r depends again on the rating p_{ait}, but is reduced by the similarity of the form of a to the form of i. w controls the weight attached to individual ascriptions as opposed to those made on the basis of similarity to something else. k and g are constants which control the speed of the decay and the primacy factors in figure 8.3.

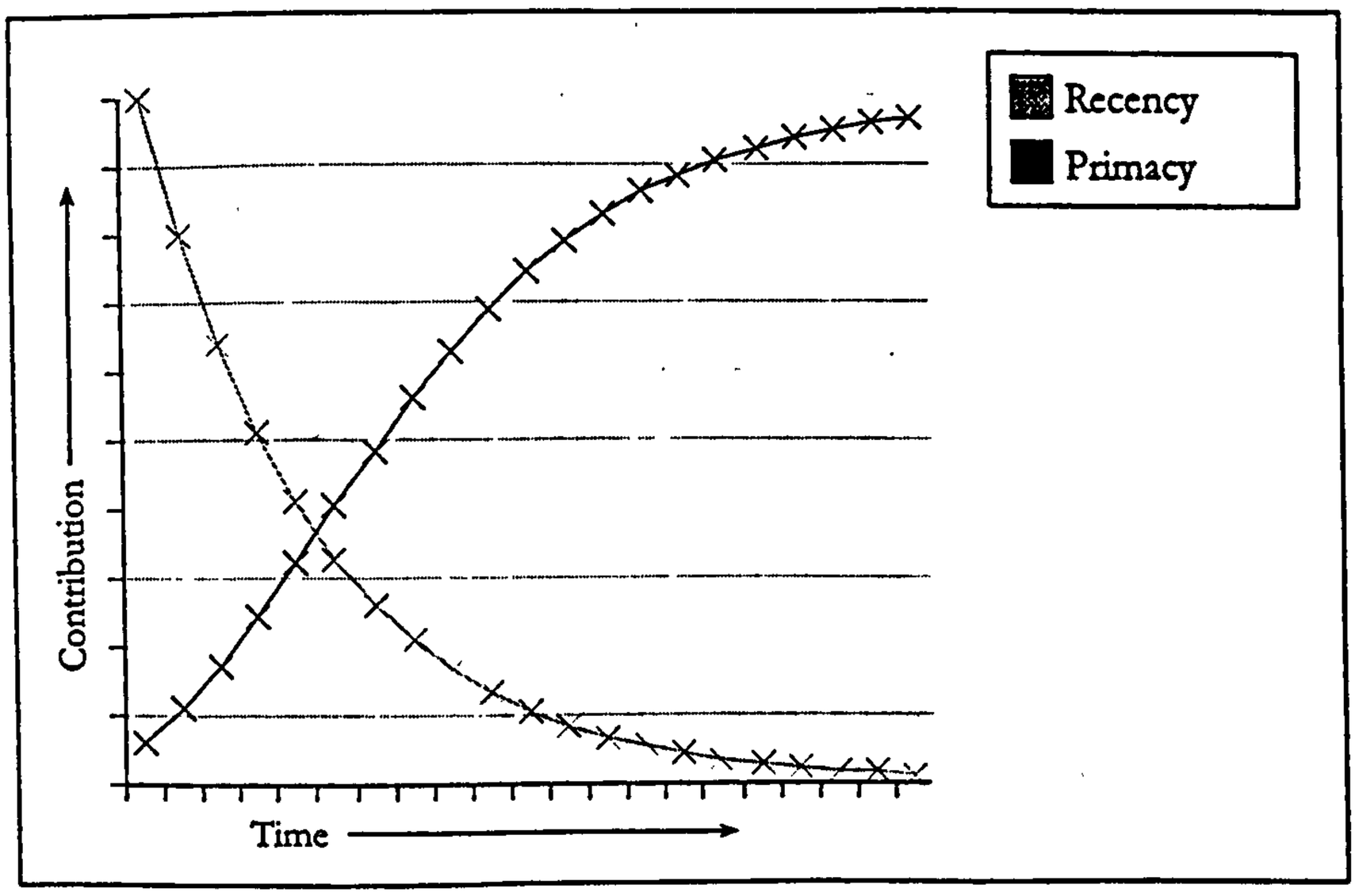

Figure 8.3. Changes in the primacy and recency of familiarity over time.

Animation

The animation factor is not modelled on any scale. There are two reasons for this. First, it is actually a very complex factor; to decide whether something is animate or not requires substantially more physical—and functional—reasoning than is needed elsewhere. Secondly, this is one of the aspects of common-sense psychology that has already been modelled in reasonable detail, by Shultz (1988; 1991). Shultz's (1991) model, like this one, aims to provide a model for that part of common-sense psychology which distinguishes things which are intentional from things which aren't, but his model rests almost completely on animation, and ignores the issues raised by the other factors of anthropomorphism.

Animation is a complex factor, and it is incredibly hard to model accurately. It is, in effect, a measure of the kind of movement of a system. In many ways, the animation factor is a measure of the movement's similarity to that of a human, although any movement at all makes a difference. It is not necessary that this movement be 'uncaused' to the extent that it can be a measure of whether the system can initiate its own movement. Many familiar objects, such as trains and aeroplanes, which are treated anthropomorphically, and which are often anthropomorphised substantially more than other inanimate objects, cannot initiate their own movement. It just seems that movement, especially certain (human-like) kinds of movement, act almost as a Lorenzian 'releaser' for anthropomorphism, under the right conditions. A particularly spectacular example of this is Penny's robot 'Petit Mal' (Penny, 1993), which has a rather unique style of movement; when people interact with it they can often be observed imitating its movement. This accords well with Meltzoff and Gopnik's (1993) theory that imitation is an important sign of common-sense psychology.

For the purposes of this model, a single feature has been assigned to the form description, which represents whether or not the form is animate. This feature is not used in the similarity measure, instead it is converted to provide a measure of animation between zero (inanimate) and one (animate). The animation factor should, in principle at least, be real-valued rather than simply discrete. In practice this representation of animation is a gross oversimplification, but, until Shultz's (1991) model of action, intention, and agency and can be combined with this one, it will do.

Structure

One of the more curious features of anthropomorphism identified in the previous chapter was the tendency that any knowledge of the structure of a system had to inhibit the anthropomorphic effect. Anthropomorphism involves ascribing a system human-like qualities, but if the observer knows all about the physical structure of the system, this can actually reduce the ascription of human-like qualities. This interacts with the causal aspect of the animation factor; that is, deciding whether a system's movement was caused by itself or by another in part depends on the observer's knowledge of the system's structure; again, this was not part of Shultz's (1991) model of animacy in common-sense psychology.

But this is only part of the story. There is a reverse effect of the structure factor. Not only can knowledge about the structure of the system tend to inhibit the ascription of mental states, a lack of knowledge—or even too much knowledge—about the structure of the system can tend to amplify the ascription of mental states. In other words, the more complex a system, the harder it is to understand in any other way, and the more likely it is that intentional explanations will be used to describe it, as Dennett's example of the chess machine shows.

I have illustrated this with a few examples. In the previous chapter, I mentioned Woolgar's (1985) point about ascribing intelligence to systems. Woolgar describes a device which bolts on to a video recorder and splices out advertisements during recording. On one level this is clearly intelligent behaviour, but as soon as we learn that it actually works by detecting a particular signal in the transmission, this changes our ascription of intelligence to the system.

Another example is Searle's (1980) Chinese Room thought experiment. Searle's thought experiment, as I mentioned in chapter 5, is really a version of the Turing test, but Searle has broken the rules by telling us how the system works inside, how it is organised structurally and functionally. This is the essence of the systems reply—that Searle has broken the rules and we must try to forget our new knowledge of the structure and see the room as a whole as an intelligent system. Collins takes the misleading nature of the Chinese Room thought experiment to depend on "our disinclination to attribute intelligence to certain kinds of mechanism" (Collins, 1990). This certainly matches the effect pretty well, but it doesn't say much about which kinds of mechanism have

this effect, or why; perhaps a close approximation can be found in anthropomorphism, but even this alone doesn't explain the positive tendency to identify with Searle-in-the-room as opposed to the negative tendency to identify with the room as a whole. This hint of negative as well as positive effects is an important one. Things—and sometimes even people if they don't seem to fit into our form of life—can 'resist' identification or anthropomorphism just as much as they can attract it.

These examples aren't intended to provide a coherent or complete argument; they are more in the nature of signposts to the source of the problem. In both these examples, there is a change in our interpretation of a system when we understand more about why it behaves the way it does—and the difference is made by additional information at a different level of explanation. So perhaps we should look in more detail at the effects of levels of explanation.

Levels as a psychological effect

In cognitive science, levels of explanation are a consistent and well-developed theme. The idea is simple and closely related to functionalism in philosophy and to the notion of a virtual machine. A level of explanation is just a way of seeing a system—but a way that abstracts away details of how that system is built. A level of explanation should be both coherent and not too large to comprehend as a whole. The standard functionalist claim is that levels of explanation can be treated with a degree of independence from each other.

Newell was one of the strongest proponents of viewing systems as a cascade of levels; his diagrammatic representation of this is shown in figure 8.4. Newell's description of levels shows how much is being read into the structure of the system by the observer: "levels are clearly abstractions, being alternative ways of describing the same system, each level ignoring some of what is specified at the level beneath" (Newell, 1990). Newell, however, didn't believe that levels are only "in the head of the observer", but that they have, to a significant extent, a genuine status in science.

Newell further described levels as being "'stronger' or 'weaker', depending on how well the behaviour of the system, as described at a level, can be predicted or explained by the structure of the system described at the same level" (Newell, 1990). But by this, Newell appeals only to scientific explanation through scientific theories; and as I showed in chapter 2 these scientific theories are often paralleled by common-sense ones, so the strength of a level also depends on whether a scientific or common-sense stance is being taken.

When we look at a system from a common-sense point of view, then again, some levels will also seem stronger than others. For example, when we think about Woolgar's video device, the additional information about the device's operation enables a better common-sense engineering description to be constructed than the previous common-sense psychological one. The engineering level increases in strength and we switch to seeing the system in this new light. But if, on the other hand, we don't know enough about the functional or physical structure of the system—or we know it in too much detail to be able to make predictions with it—we stay with the intentional description.

Searle's example is similar, with one extra gloss. One of the components in the apparently functional description is a human—Searle-in-the-room himself. If anything in the new description of the Chinese Room corresponds to a strong level as far as common-sense psychology is concerned, it is Searle-in-the-room; it is too easy to identify with Searle-in-the-room—that's most of

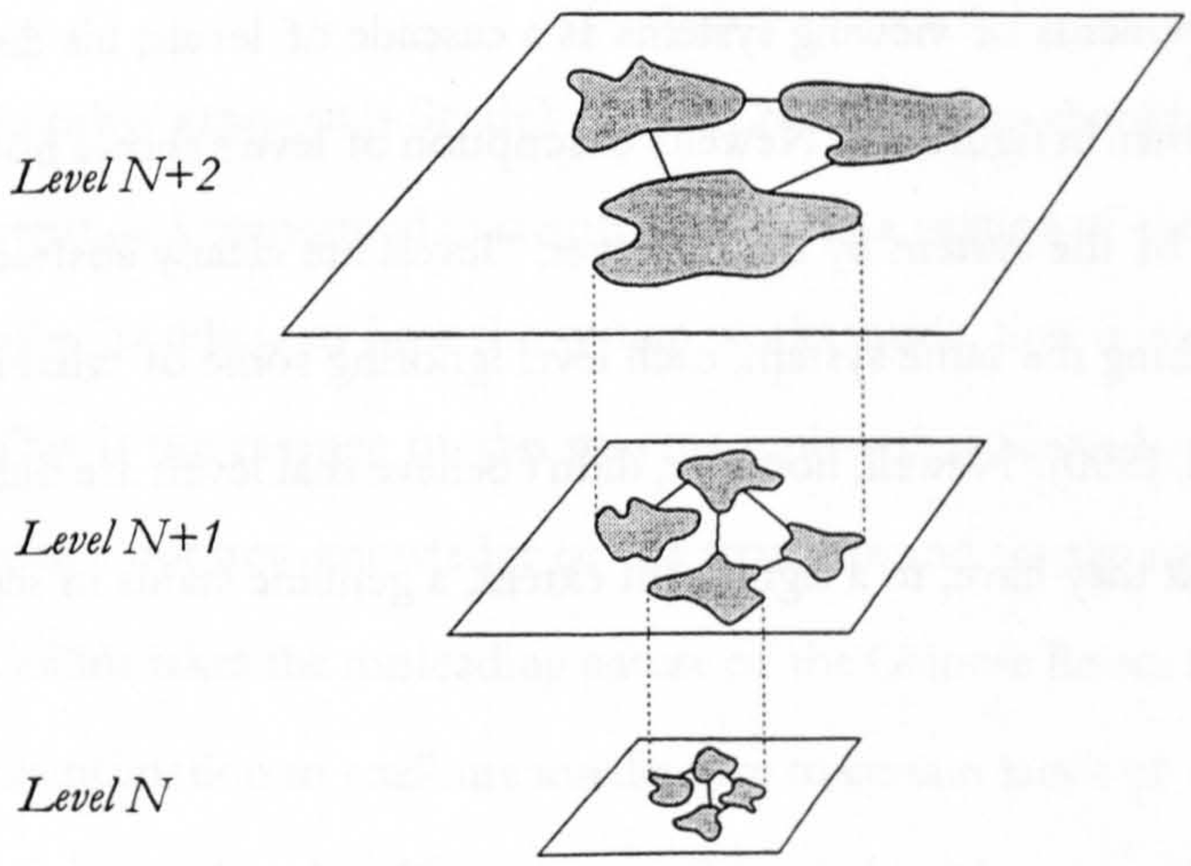

Figure 8.4. Levels of explanation (after Newell, 1990)

the problem. Unfortunately, the rules of the game have been written so that Searle-in-the-room doesn't count, because he's just playing the role of an automated system. So we're stranded between the levels, and we can't make sense of the system as a whole.

This view of levels is neither traditionally functionalist nor conventional, nor is it intended to be. The point I am making is that while levels might not be 'in the head of the observer' as far as scientific theories are concerned, as far as common-sense theories are concerned, in great measure they really are. And even in science our common-sense judgement can significantly influence our scientific interpretation of the levels involved, as I'll discuss in chapter 14. This doesn't invalidate the principle of levels; indeed it offers a new insight to them through the principles of common sense.

Perhaps the best argument for this view of levels is by attacking the alternative, the hypothesis that levels are something real; something that really exists. There are several possible criticisms of levels. The first, and perhaps simplest, is the evolutionary criticism—although levels of folding can happen evolutionarily, in general, strong levels of the kind usually taken in the virtual machine sense do not seem to arise in evolution. In practice, evolved systems are far more chaotically organised than that. Functionally disparate levels do not generally evolve any more than functionally disparate modules of any other kind. A second criticism is that levels are closely associated with the idea of functionalism, and the whole idea of functional equivalence is on rather shaky ground when anthropomorphism is taken into account. This is because in anthropomorphism, it is possible for functionally identical systems to be seen, by an observer, in completely different ways. An excellent example of this is Haugeland's (1980) neural reply to Searle's Chinese Room, which I'll come back to in chapter 11.

Note that I am not arguing that levels aren't real, just that, strictly speaking, they are products of our intuitions and our common-sense psychology, not simply of the actual structure or behaviour of the system. I will discuss the close connection between intuitions and levels of description more fully in one area, that of consciousness, in chapter 14. For now, I am simply suggesting that

the intuitive appeal of different levels of description is subject to the principles of common-sense psychology, and, therefore, levels are also subject to all the anthropomorphic effects and individual differences that accompany them.

The structure factor to anthropomorphism is very important here, but it is not particularly obvious how it works. Again, there is not much direct empirical evidence. I'll return to its significance later, both in the model of the Chinese Room in chapters 11 and 12, and in chapter 14, where I'll show that it does seem to have important and subtle effects on thought experiments. In practice, the model of the structure factor in this thesis is still very vague; a lot more work needs to be done to get a better idea of the kinds, qualities, and quantities of physical and functional knowledge that have this effect. For example, the feature attribute model of similarity metrics used in the models seems to be implausible, given the current evidence (Lakoff, 1987); yet on the other hand, the similarity measure cannot be completely learned, because it seems to have effects even on neonates.

Returning to the structure factor for anthropomorphism, we can model the effect of the structure factor, to a first approximation, by the following formula:

$$str(a, i) = \frac{k}{n^2}\left(\sum_{j=1}^{n} sim(f_a, b_j)\right)$$

Here the structure factor depends inversely on n, the number of elements which the agent a believes to be inside the form i, and on the sum of the similarities of these elements to a. The effect of the structure factor, then, is strongest when there are small number of elements inside i, and when these elements are similar to the ascriber.

Context

Again, context is not especially easy to model in a way that is useful. Context changes according to the situation in a way that is, currently at least, not very predictable. Part of it, undoubtedly, is somewhat like a spreading activation; in that once a situation is created where anthropomorphism works, this seems to increase the tendency toward anthropomorphism. But for now, in this model, the context is represented as a fixed number, again between zero and one, which conveys some

idea of the current tendency toward anthropomorphism. Increasing this number will, for a given scenario, tend to promote anthropomorphism of objects which would probably otherwise be classified as physical rather than mental.

In practice, this is an oversimplification. There can be both long term and short term changes in the context. Pretence (Leslie, 1987), for example, can make a temporary change in the patterns of anthropomorphism; and even Baron-Cohen *et al.*'s (1985) false belief test depends on the child being able to pretend that a doll represents a person for the duration of the test. This is very different from the long term changes in context discussed in chapter 5 in the context of the Turing test—changes, for example, in the generally accepted meanings of words.

Medium

The effects of the medium of interaction are not substantially addressed in this model, that is: there is no computational model of the effects of the modalities involved in an interaction medium. Although this might be possible in theory; in practice the effect of the medium is very subtle and can be hard to distinguish from the effects of the form and the familiarity.

The paradigm example of the medium effect is shown in the Turing test. In the standard version of the Turing test, all interaction is carried out through the medium of a teletype link. Turing's argument for this was that linguistic interaction alone is enough to determine whether or not to ascribe intelligence to a system. While this is probably true, there is also some evidence (e.g. from Hofstadter's, 1985, description of a reversed Turing test, discussed in chapter 5) that using a machine-like mode of interaction will create a disposition to ascribe machine-like mentality—even when the system at the other end is a person and is behaving completely normally—without pretending to be a machine.

This seems to be rather contradicted by the example of electronic mail. People, on a day to day basis, have no problem ascribing mentality to people they correspond with by electronic mail. They do not, of course, *expect* a correspondent to be a computer—so they aren't trying to make the same kinds of judgement as they perhaps would in the Turing test, and in this sense, there is a

connection with the subjects' failure to find fault with their random counselling advice in Garfinkel's (1967) experiment, described in chapter 5. Even so, the effects of the medium seem both to be contextual and to depend substantially on the ascriber's previous experience. This is a factor that needs to be explored far more in experimental situations, like that of the Turing test, before adequate models can be constructed. It seems wise, therefore, to leave the medium constant for the sake of these models.

Predicate

Finally, a decisive factor in the selection of the stance may be provided simply by framing the question in a particular way. If we take the example of a thermostat, it is clear, from McCarthy and others, that it is possible to take both a physical and an intentional stance to it under different circumstances. The selection between them can be determined not only by all the other factors presented here, but simply by asking the question 'does the thermostat think it is too cold?' rather than 'is the thermostat switched on?' Both questions address the same semantic content, but the stances that are involved could be very different.

In practice, in some cases, there will be a dissonance between the stance suggested by the question predicate and the form and behaviour of the system; when this happens, there will be a clash of intuitions which might result in not being able to answer the question because, in effect, no stance can be taken. McCarthy's (1979; 1983) thermostat example shows this kind of dissonance, but imagine this: instead of asking 'is this thermostat switched on?' ask 'is this person switched on?' The change in form in the second question pushes it to a metaphorical interpretation rather than a literal one, because the physical emphasis of the question is not compatible with our physical knowledge about people. The question changes because of this dissonance effect.

Bringing the measures together

All of these different factors are modelled as measures with a value between zero and one, but before these can be used they must be combined into a single value; a value which can be used to determine which stance should be selected. For now, we'll make the simplifying assumption that there is no significant interaction between these factors. This assumption is certainly false, but at present the factors are not yet well enough defined to be able to decide where and how strong these interactions actually are. Assuming that the factors don't interact will make the model tractable in the first instance, and leaving refinement of our understanding of the probable interactions to future versions.

Being dispositional, each stance can be modelled as a set of behaviours together with a function which reflects how strong the disposition to use those behaviours currently is. The model implements this behaviour by attaching an 'energy' value to each stance. When the time comes to select a stance, the stance with the lowest energy value is chosen and, if it is different from the current stance, a switch is made.[1] Each stance's behaviours can be represented conventionally using a set of rules, for example, to model a theory or simulation; this is sufficient for the descriptive account we want. The energy for a stance is a function of the values attached to all the dispositions for this stance. Selecting a stance, then, corresponds to choosing the stance with the lowest energy. In principle, this could be achieved by any optimisation function, but for the sake of simplicity, an absolute minimisation is used in the model. This use of an energy function requires a bit of explanation, because it actually reflects a subtle slide into a new modelling framework, but the energy function shouldn't be regarded as anything to be scared of. On the contrary, it is a fundamental principle behind many physical and psychological models, such as those of neural networks (Amit, 1989).

[1] Using the lowest energy, rather than the highest, might seem counterintuitive, but this is conventional for energy function systems like this, because the underlying principles of energy systems are those of thermodynamics. These principles underpin many different kinds of optimisation, not just those of physical systems; Amit (1989), for example, shows that these thermodynamic principles provide a useful foundation for many different aspects of neurophysiological behaviour. Note that because we select the stance with the lowest energy, a disposition towards a stance will reduce its energy value, and a disposition against it will increase its energy value.

Figure 8.5 shows perhaps the most common way of thinking about the behaviour of an energy function system like this. Imagine the space of possible states of a system as a lumpy landscape, over which the system moves a bit like a marble which is continually being jiggled. Over time, the marble will tend to settle into the lowest hollow in the landscape—the lowest energy point for the system. Note that although the lowest energy point for the system is at a particular point, the system itself is dispositional—it operates as a tendency for the system to move towards that particular point rather than the point itself acting on the behaviour of the system. The hollows don't 'attract' the marble, but the shape of the system is such that there are hollows, and the marble falls down the hills into them.

Relating this model to the process of selecting the right stance, we can imagine that the landscape is not fixed, but changes its topology from moment to moment depending on the dispositions. The lowest energy point for the system, therefore, can and will move dynamically depending on the stance expected at any moment in time. As each stance is associated with an energy function, the overall system will have a disposition to move towards the stance with the lowest energy value.

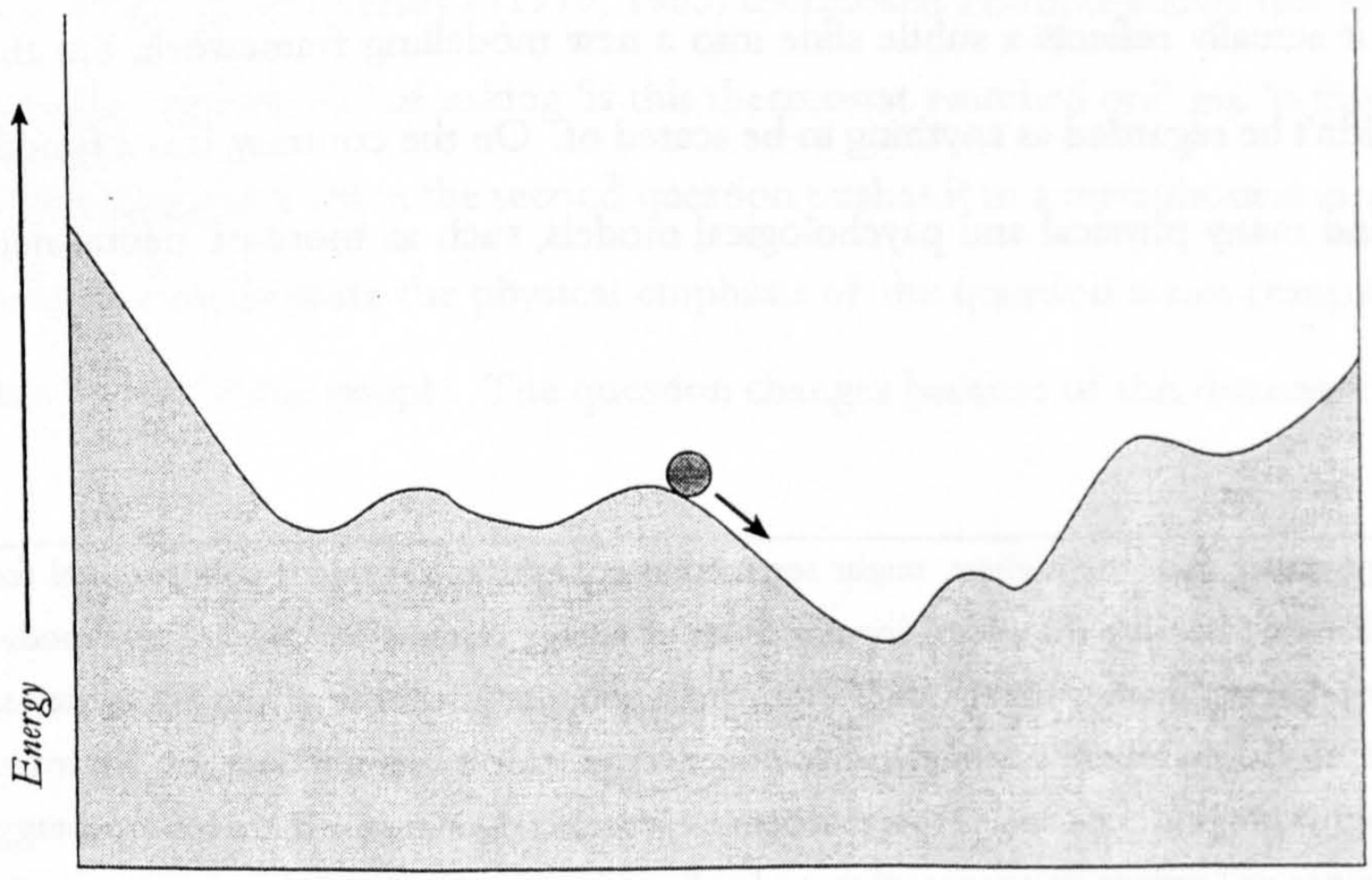

Figure 8.5. An example energy landscape

Summary

I have now developed and clarified the theory of the previous chapter considerably. I have discussed the different stances, the relationships between them, and the factors that influence the selection of one stance over another in more detail. In the next part of the thesis, I'll continue this elaboration, and develop this paper theory into a computational model of the processes involved in selecting between—and using—the intentional and physical stances. Then, to show the model to full effect, I'll use this model in two studies; first, to compare different theories of common-sense psychology in Baron-Cohen *et al.*'s (1985) false belief test, and secondly, to show the effects of intuition on Searle's (1980) Chinese Room thought experiment.

BLANK IN ORIGINAL

Part Three

Modelling common-sense psychology

BLANK IN ORIGINAL

Chapter 9
Models of common-sense psychology

Introduction

In this third part of the thesis, I'll show how we can use the theory of common-sense psychology developed in chapters 6, 7, and 8 to build a computational model of the processes of ascription. This model will illustrate a simple environment in which we can represent people ascribing mental states to each other and, through anthropomorphism, to objects as well. As part of this, it will provide a simple model of the factors involved in anthropomorphism, of their influence over which stance is selected, and basic models of both a physical and an intentional stance.

This chapter will lay out the general structure of the modelling environment, and show how the models of anthropomorphism and the different stances relate to the more theoretical analysis of these issues in the previous part of the thesis. Following on from this, the other chapters in this part of the thesis will show how this general modelling environment can be used to study the behaviours characteristic of common-sense psychology more fully.

I'm going to present two different studies through this model. The first of these is Baron-Cohen *et al.*'s (1985) modified false belief task, described in chapter 3. This is a good first model to work with because it is fairly straightforward, requires only a little physical reasoning, and, especially importantly, it shows how the approach can contribute to research by allowing subtly different theories of common-sense psychology to be compared. I'll describe this study fully in the next chapter.

The second study is more complex. It is a model of the ascription processes involved in Searle's (1980) Chinese Room thought experiment, described in chapter 5. At first, this might not seem like a particularly good arena for this kind of model, but in practice it exercises several factors (e.g.

familiarity and structure) which are not particularly relevant to the false belief test. The Chinese Room thought experiment is derived from and related to the Turing test, so this second study will indirectly show some of the effects of these factors on the Turing test, and on the ascription of mental states to non-human systems in general. I'll describe the theory behind this second study in chapter 11, and the model for it in chapter 12. In this chapter, though, I'll set the scene and show the common ground to both these studies, namely how the modelling environment implements the general foundation of anthropomorphism in common-sense psychology that is used both in the false belief test and in the Chinese Room.

Methodological issues

Before launching into the models proper, it is worth reviewing the methodological principles underlying the approach that I've taken—that of computational modelling. This is a controversial issue in its own right, for philosophical reasons as well as methodological ones, but it is right to be clear on these matters. Proper methodological analysis of computational modelling is fraught with dangers, as there are tangles between the theoretical and methodological issues which can contaminate the methodology. For instance, although I am using the methodology of computational psychology, if I take computational psychology as drawing "on the concepts of computer science in formulating theories about what the mind is" (Boden, 1988) it says something about the theory as well as about the methodology. In fact, the theory I am developing in this thesis actually throws some doubt on this use of concepts from computer science, despite its use of the same method. It is for this reason that I have adopted the distinction between engineering and descriptive accounts made by Clark (1988), and discussed in chapter 2. The problems with concepts from computer science derive from the structure factor in anthropomorphism, which seems to show that there could be a genuine psychological tendency for our intuitions to betray us when we think about certain kinds of mechanism underlying a cognitive model—all theories are not the same, at least as far as common-sense psychology is concerned. I will return to these methodological issues in chapter 14.

Of course, restricting the model to a descriptive project rather than engineering project weakens it considerably; one of the apparent virtues of using computational models is the implicit strength of the accounts involved. But the account is not weakened beyond usefulness; it just becomes a model of the same strength as any other psychological model expressed on paper. The other virtues of computational models—their precision and relative unambiguity—is retained, and these virtues hold for a purely descriptive account just as much as they would for an engineering account.

McCarthy's 'situation calculus'

Before we can build models adequately, we need a representation language which is strong enough to be able to do the physical and psychological reasoning needed. In practice, at least the psychological parts of this model require the ability to reason about different contexts (after all, this is at the heart of the false belief test) so something a bit like modal logic is going to be needed, as shown in chapter 4. The model borrows this from McCarthy's (McCarthy & Hayes, 1969) 'situation calculus', where the effects of an event are described as a consequence relation between one state and another. At the core of McCarthy's calculus is a special function, *result*, which can represent the effects of an action on a situation, by returning a new, modified, situation. The function *result(p, σ, s)*, where *p* is a person, *σ* is an action, and *s* is a situation, has a value which is a new situation representing the effects of *p* doing *σ* in *s*. Thus, for example:

$$inside(marble, X, s) \wedge \neg inside(marble, box, s) \Rightarrow inside(marble, box, t) \wedge \neg inside(marble, X, t)$$

where $t = result(alison, putin(marble, box), s)$

This says that if *marble* is inside something that isn't *box* in a situation *s*, then the effect of *alison* putting *marble* in *box* in situation *s*, is a new situation such that *marble* is no longer where it was (in *X*) but is now inside *box*.

The situation calculus is both more powerful and more complicated than this implies, but this subset of it is sufficient for the purposes of this thesis, and furthermore, it doesn't need the heavy inference machinery that a complete modal logic would. There is, however, a connection between

modal logics and the situation calculus, in that modal logics can be reconstructed within the complete situation calculus (McCarthy & Hayes, 1969). The situation calculus, then, is strong enough for this thesis, relatively easy to use within a computational environment, and yet it retains the referential properties of the modal logics presented in chapter 4, referential opacity for example.

The modelling environment

Before moving on to the psychological parts of the model, which are at the heart of this thesis, I will discuss some of the principles behind its implementation—the implementation itself is in Appendices C and D. The model is implemented in a Prolog-like language embedded in Common Lisp. There are two principal deviations from standard Prolog—apart from the Lisp-like syntax. First, this language allows the use of many separate databases. The model uses this to represent the different objects. Apart from a few global rules, each object in the model has rules in its own separate database, and can reason with, assert to, and retract from, that database without affecting other objects. The different stances also have their own databases; so selecting a stance is equivalent to temporarily selecting a database, and while that stance is selected all reasoning is subject to the rules belonging to that stance, as well as to the rules which are global or which belong to the object.

This technique is also used to highlight the differences between models. By joining databases together, it is possible to build, for instance, a version of the intentional stance that models the simulation theory, but which inherits from a more general basic version of the intentional stance. Then, if no rules are found in the specific database for the simulation theory's intentional stance, the interpreter will pick up the more general rules belonging to the general intentional stance. This approach emphasises the rules that are different between the models.

The second change from standard Prolog is purely notational—in a clause's head, input variables are labelled with a *?* (question mark) character, and output variables with a ^ (caret) character. Within a clause, variables labelled *?situation* and *^situation* refer to the same variable, but the added caret helps to show that values are written into that variable, rather than read from it.

The standard model: Leslie's 'decoupler' theory of mind mechanism

To compare the different theories, I'll begin by introducing and describing Leslie's 'decoupler' model for common-sense psychology. Although common-sense psychology is hugely complex, and can only be modelled in the most sketchy form, I'll show how Leslie's theory can be implemented as a computational model. Then, alternative theories of common-sense psychology can be represented as variations on this basic theme, and we can draw some conclusions about the similarities and differences between the theories within this modelling framework.

Leslie's model is shown in the diagram in figure 9.1. At the heart of Leslie's model is a manipulator that is capable of pretence—of decoupling beliefs from one context and applying them in another. It is this that makes reasoning about false beliefs possible, because this decoupling mechanism separates a child's own beliefs into a different context from the same child's model of other peoples' beliefs.

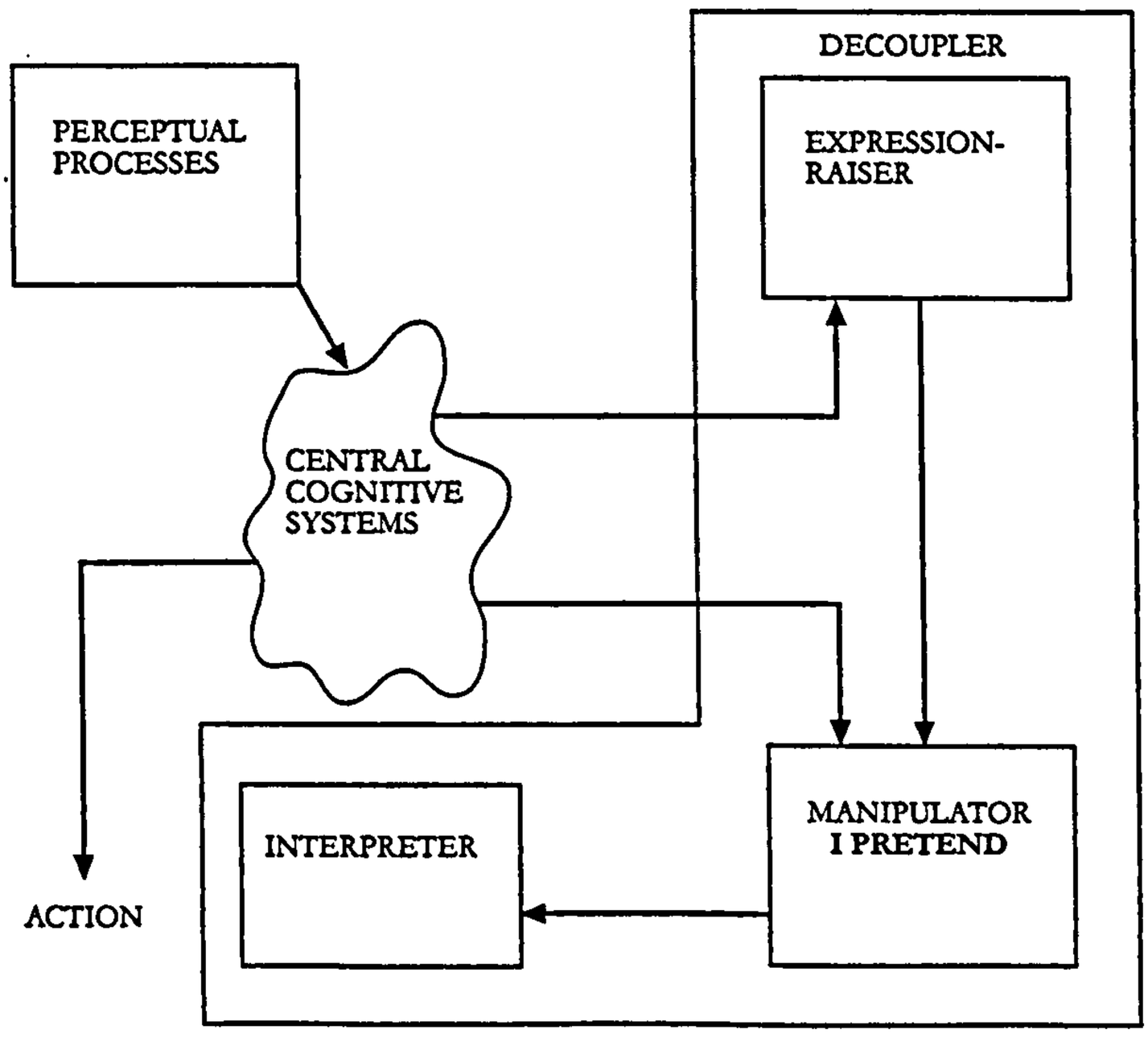

Figure 9.1. Leslie's 'decoupler' model (after Leslie, 1987)

Given this theory of common-sense psychology, we can now turn to the model, and show how Leslie's theory can be represented computationally. The model comprises a number of separate modules. These are: a physical environment model, general rules for people, a module for selecting between stances, basic dispositional factors, and basic physical and intentional stances.

The physical environment model

The first part of the modelling environment is a physical environment model, which implements an event-driven simulation environment. As objects are moved physically from one place to another, events are generated, and these events will be received and handled by all objects equipped with sufficient perceptual apparatus to be aware of them.

In practice, the physical environment model uses only the principles of containment; that is: object *A* can (in principle) see object *B* if and only if they are contained by the same object. When an object *A* moves into a container *C*, all objects already inside *C* become aware of it, as do all the objects outside *C*, which is where *A* moved from. In addition, as *A* enters *C*, it becomes aware of the placements of all the other objects inside object *C*. This model could, of course, be enhanced by a more complex physical simulation in the manner of Davis (1988) involving, for example, objects of different sizes, shapes, and positions, lines of sight, and so on, without affecting the principles behind the model.

This physical environment model is implemented by the rules in figure 9.2. The physical environment model is used whenever an object is moved in the physical environment; this move is passed as a query to the *tell* relation. The *tell* rules do two things; first, for any move they manage the world model by asserting and retracting *inside* facts to keep everything complete and consistent. These rules ensure that the *inside* relation will always be a many-to-one relation—never a many-to-many—so that every object is only in one place at any one time.

The second task of the *tell* rules is more complicated; this uses these *inside* relations to decide which other objects in the world model will be aware of the change in the physical world and notifies them of it, by sending them an event through the *notify* relation; this is where the princi-

ples of containment described earlier come in. For example, if an object enters a room, it will receive events which correspond to its seeing all the objects already within that room, each in their proper place. These events are generated irrespective of whether or not the object has any perceptual capabilities, and irrespective of whether or not it will ever be aware of the event; I'll come back later to the effect of these events on different objects' perceptual systems.

```
;;; If an object is put into a container, we find out where it was before, and assert that it has moved
;;; from there to the new container. We also notify all other objects in both locations of this move,
;;; and the object itself of everything it can see in its new location.

((tell (put-in ?object ?container)) :-
 (inside ?object ?outer-container)
 (retract ((inside ?object ?outer-container)))
 (asserta ((inside ?object ?container)))
 (notify-all-inside ?container (perceived (put-in ?object ?container)))
 (notify-all-inside ?outer-container (perceived (put-in ?object ?container)))
 (notify-object-places ?object ?container))

;;; If an object is taken out of a container, we find out where it was before, and assert that it has
;;; moved from there to the new container. We also notify all other objects in both locations of this
;;; move, and the object itself of everything it can see in its new location.

((tell (take-out ?object ?container)) :-
 (inside ?container ?outer-container)
 (retract ((inside ?object ?container)))
 (asserta ((inside ?object ?outer-container)))
 (notify-all-inside ?container (perceived (take-out ?object ?container)))
 (notify-all-inside ?outer-container (perceived (take-out ?object ?container)))
 (notify-object-places ?object ?outer-container))

;;; These rules notify all objects inside ?container of an event ?event.

((notify-all-inside ?container ?event) :-
 (bagof ?object (inside ?object ?container) ?objects)
 (notify-all ?objects ?event))

((notify-all () ?event))
((notify-all (?object . ?rest) ?event) :-
 (notify ?object ?event)
 (notify-all ?rest ?event))

;;; These rules notify an object ?object of the places of all the objects inside the ?container.

((notify-object-places ?object ?container) :-
 (bagof ?other-object (inside ?other-object ?container) ?objects)
 (notify-place-all ?object ?container ?objects))

((notify-place-all ?object ?container ()))
((notify-place-all ?object ?container (?other-object . ?rest)) :-
 (notify ?object (perceived (place ?other-object ?container)))
 (notify-place-all ?object ?container ?rest))
```

Figure 9.2. Rules for the physical environment model

Note that this physical environment model doesn't use any principles from the situation calculus. This is as it should be—the physical environment model is intended to be a description of 'reality', even though this is a reality that might never be known to any person, animal, or object represented in the modelling environment. Because it is a model of reality, there can never be more than one situation and the changes can be applied to that situation directly; and as a model of reality, it cannot necessarily be directly perceived by any of the agents that live within the environment that it provides.

Finally, the physical environment model provides the *sees* relation, which is a primitive that holds for all the objects that are physically inside the same container as this object; again, this doesn't depend at all on the objects' perceptual capabilities, only on the physical layout. A more sophisticated *sees* relation could be provided if required, perhaps using some of the principles of Davis (1988) to use line of sight in determining when one object can see another. It should be noted, however, that there is an important distinction between *sees* in the physical environment model, and a common-sense psychological notion of *sees*, which is ascribed, and which might be expressed, for example, by '*A* can see that *B* can see *C*'. In this model, I am largely ignoring this distinction, partly because this seems to be a notion that changes as children develop (Carey, 1985) and is therefore not constant, but also because within the limits of this model, the two notions behave identically.

Modelling people

When an object is going to be notified about an event, the object's clause database is selected, and the expression *(handle-event ?event)* is queried against that database, where *?event* is a variable instantiated to the event in question. This corresponds to the event happening in the world, but whether or not the object responds in any way depends upon the object's own clause database. This is where the individual object's perceptual systems are modelled. Objects without perceptual systems, simple physical objects, for example, won't have any rules to handle the event, so it ends up just being ignored.

The rules in figure 9.3 show how the model of a person latches into these events by providing rules for the *handle-event* relation. The effect of these is to apply the *object-event* relation for all the objects that match the *sees* relation. This roughly corresponds to the informal, theory of mind style rule 'if I can see something happen, and I can see you, then you can see it happen too.'

As well as handling events, people can also handle questions, through the rules in figure 9.3. This works in a very similar way, except that the question should already relate to a nominated object; that is, questions can take the form of 'where does Sally believe the marble is?' rather than 'who believes that the marble is in the box?' The model for questions also allows an answer to be returned in the *?response* variable.

The *object-event* rule is at the core of the model for a person. Each person has their own situation, which is retrieved by the *get-situation* relation, and stored by the *put-situation* relation. This situation is used to keep all that person's notional worlds, for all the objects that they are aware of and

```
;;; Handle events in a failure-driven loop.  For each object that this object sees, use the object-event
;;; primitive to ask the object to set up the right stance to ?someone for this ?event.

;;; Rule handle-event; called by the physical model
((handle-event (perceived ?event)) :-
 (self ?self)
 (sees ?self ?someone)
 (object-event ?response (perceived ?someone ?event))
 (fail))
((handle-event ?event))

;;; Questions are handled similarly, except that the object is already provided in the question, and
;;; we return an answer response in the ?response variable.

;;; Rule handle-question; called when you want to ask somebody a question
((handle-question ^response ?question) :-
 (object-event ?response ?question))

;;; Object events are handled by a rule that gets the current situation with the get-situation
;;; primitive, and then uses the in-stance primitive to activate the appropriate stance to the object
;;; ?someone for the predicate ?action.  The response is returned in ?response.  The new situation is
;;; stored by the put-situation primitive.  The get-situation and put-situation primitives are shown in
;;; more detail in the file basic.component, in Appendix C.

((object-event ^response (?action ?someone ?event)) :-
 (get-situation ?situation)
 (in-stance ?someone ?action
   (result ?response ?someone (?action ?someone ?event)
           ?situation ?new-situation))
 (put-situation ?new-situation))
```

Figure 9.3. General rules for modelling people

that they can take the intentional stance to. The rule, then, gets the person's current situation, applies the *result* function to it—after selecting the appropriate stance—to get a new situation for that person, and then finally stores the new situation for the future. The *result* function is the *result* function of McCarthy's situation calculus, discussed earlier. McCarthy's function took an agent, an action, and a situation, and returned a new situation. The *result* function in this model is identical except that it returns one additional response value; this is the value used to return the answer to a question.

Selecting the right stance

Stances are selected by the *in-stance* primitive (defined in the file *model-primitives.def* in Appendix D), which takes a target object, a predicate, and a query form. This then calculates and combines all the available dispositional factors for all the available stances to the given object for the given predicate, and selects the stance with the lowest total dispositional energy. This corresponds to movement on an energy landscape, as illustrated in figure 8.5. Then, with the stance and its database temporarily selected, the query will be passed to the interpreter; any of the stance's rules can then participate in the reasoning process. When the query is complete, the selected stance and its corresponding database are restored to their original status—their rules will no longer be available until the stance is selected again.

There is a bit more to taking the right stance than this. With people, for example, even if the right stance is not selected at first, they will quickly adjust and carry on, although there will often be a dissonance effect, representing the conflict between people's expectations and the reality of the system's behaviour. This 'anthropomorphic dissonance' could be represented by a kind of backtracking, so that when a prediction doesn't work well, the next best stance is selected, and the ascriber will learn from this dissonance, and may use a different stance next time they try to predict the same system's behaviour. This dissonance is an important effect, and one I will return to in chapter 14, where I'll suggest that it plays an especially important role in psychologists' understanding of cognitive models. It is, however, an effect that needs considerably more re-

search before we can build an accurate model of it. In the first instance a simple model of the dispositional factors is enough, and we can assume that the ascriber always has dispositions which select an appropriate stance.

The basic dispositional factors

So far, I've described the general structure behind the model's selection between different candidate stances, but this has left out any detailed description of the model's interpretation of the factors that influence this selection. In part this is deliberate; in chapter 12, I'll use this to show that different individuals may have subtly different factors, and that this can significantly affect their ascription of mentality. In the meantime, I'll briefly summarise the model's general representation of these different factors and their combination; following on from the more complete descriptions of these factors in the previous chapter.

Modelling similarity. (See the file *similarity.factor* in Appendix C). To represent similarity, I have used a model derived from numerical taxonomy, which combines the product-moment correlation coefficient (for real-valued features) with the Rogers-Tanimoto coefficient (for attribute features). All the objects in the model are described for a set of form descriptions, which provide a rough description of the physical form of an object as a set of feature attributes. Every object in the model is required to have some physical form, and for the correlation coefficient in particular, there must be at least three features which have real values. In practice, in numerical taxonomy, the number of features to be taken into account is usually greater than 60. The standard form descriptions used within the model are shown in figure 9.4, and represented in the file *forms.component* in Appendix C.

Modelling familiarity. (See the file *familiarity.factor* in Appendix C). Familiarity, as I discussed in the previous chapter, shows aspects of learning, but it also links to the same similarity metric, because familiarity can be transferred from one form to another, similar, form by analogy. I've already shown the formula that is used for calculating the familiarity factor in the previous chapter. It depends on remembering successes and failures from previous attempts to take this stance to objects of a given form. Ratings are remembered by the *add-rating* rule, which for an individual

Person

Character	Value
Has head?	Yes
Has limbs?	Yes
Solid?	Yes
Animated?	Yes
Width	0.25
Height	1.0
Depth	0.125

Doll

Character	Value
Has head?	Yes
Has limbs?	Yes
Solid?	Yes
Animated?	No
Width	0.05
Height	0.2
Depth	0.025

Robot

Character	Value
Has head?	Yes
Has limbs?	Yes
Solid?	Yes
Animated?	No
Width	0.25
Height	1.0
Depth	0.125

Room

Character	Value
Has head?	No
Has limbs?	No
Solid?	No
Animated?	No
Width	4.0
Height	3.0
Depth	4.0

Cupboard

Character	Value
Has head?	No
Has limbs?	No
Solid?	No
Animated?	No
Width	1.0
Height	2.0
Depth	0.5

Marble

Character	Value
Has head?	No
Has limbs?	No
Solid?	Yes
Animated?	No
Width	0.01
Height	0.01
Depth	0.01

Neuron

Character	Value
Has head?	No
Has limbs?	No
Solid?	Yes
Animated?	No
Width	0.00001
Height	0.00001
Depth	0.00001

Box

Character	Value
Has head?	No
Has limbs?	No
Solid?	No
Animated?	No
Width	0.03
Height	0.02
Depth	0.02

Basket

Character	Value
Has head?	No
Has limbs?	No
Solid?	No
Animated?	No
Width	0.03
Height	0.02
Depth	0.02

Chocolate

Character	Value
Has head?	No
Has limbs?	No
Solid?	Yes
Animated?	No
Width	0.07
Height	0.02
Depth	0.005

Figure 9.4. Physical form descriptions used for modelling similarity.

of a given form, stores a rating for a particular stance (*add-rating* is shown complete in file *familiarity.factor* in Appendix C). This rating will be between −1 and +1, and represents the success of the prediction given that stance. If the prediction was 100% accurate, therefore, the rating will remembered as +1; if 100% inaccurate as −1. These ratings are then summed and normalised with primacy and recency effects to return a value between 0 and +1 for each stance.

Modelling animation. (See the file *animation.factor* in Appendix C). Animation is not represented in any substantial way in this version of the model; the current representation of this factor is a dramatic oversimplification. For now, animation is simply represented as a form attribute forming part of the object's form description, although this attribute is not used in the similarity factor calculations. Instead, this form attribute is just used to hold a value indicating whether or not the form is animated. A more complete model of this factor could be provided by Shultz's (1991) model—although again Shultz's agency factor was a switch-like, on or off, value, and it is not yet clear whether the animation factor really is discrete like this.

Modelling structure. (See the file *structure.factor* in Appendix C). Structure, like familiarity, is sensitive to the current state of the ascriber, but this time, instead of the kind of learned experience that influences familiarity, it is the ascriber's knowledge of the physical structure of the object that dominates the strength of the dispositional factor. This factor is special in that it negatively affects the selection of the intentional stance; that is, knowing more about the physical structure of the object we are taking a stance to will reduce the dispositional tendency to take the intentional stance. In the model of structure I have combined two interlocking effects, those of scale and of similarity. The scale effect works inversely—the more objects there are inside a system, the greater the tendency to take the intentional stance. The idea of this is that models involving billions of components are, paradoxically, easier to anthropomorphise than those involving only a few. I'll return to this effect in chapter 14. The second effect of similarity is perhaps the one most evident in Searle's Chinese Room—and it will be demonstrated in chapter 12; the idea is that the more human-like a component of the system is, the stronger the disposition against taking the intentional stance to the system as a whole. That is; Searle-in-the-room's human form makes it harder for us to take the intentional stance to the room than it would if we had just had a computer there

in the first place. The effects of structure, though, are both complex and curious, and while this model seems to make a good first attempt at this dispositional factor, there is a lot more research that needs to be done on it.

Modelling context. Context, like animation, is not modelled fully; it is too complicated and subtle for that. Instead, I have taken the assumption that, during any single study with the model, the context will be constant; it is, therefore, represented as a constant numerical value.

Modelling predicate. (See the file *predicate.factor* in Appendix C). The predicate is modelled in a very naive way. The model recognises a number of standard predicates, and simply assigns each a rating with respect to every possible stance. The predicate is derived from the prediction that is in hand, and is passed as a parameter to the *in-stance* relation. For example, predictions which depend on physical reasoning, like whether or not an object will be inside the box after it is placed there, tend to promote the physical stance; predictions which depend on psychological reasoning, like whether or not an object believes that the marble is inside the box, tend to promote the intentional stance. Of course, these need not be exclusive; the same predicate can promote both stances.

Combining the factors. All these factors are designed to separately return a value between 0 and +1, but this isn't sufficient to describe the complete dispositional energy function assigned by combining these dispositional factors. There are two reasons for this: first, the factors need not all have the same, or even nearly the same, weight; and second, some factors may affect the selection positively and others negatively, and therefore these positive weights may need to be converted to negative ones. This conversion is modelled by allowing each of these weights to be scaled, reflected, and translated before they are combined into a complete dispositional weight for a given stance. Once again, these scaling and translation factors depend on the individual, and may therefore be subject to some minor variation between individuals. We'll see this in action in chapter 12. Tables 9.1 and 9.2 show the default scaling and translation factors for the model, and their effects on the scale limits of 0 and +1, for the dispositions affecting the physical stance and the intentional stance respectively. Figures 9.5 and 9.6 show how these dispositions are implemented in the modelling environment.

Factor	Scale	Translation	0 maps to	+1 maps to
Context	0	0.5	0.5	0.5
Predicate	−3	1.5	1.5	−1.5

Table 9.1. Scaling and translation factors for the physical stance

Factor	Scale	Translation	0 maps to	+1 maps to
Similarity	−2	1	1	−1
Familiarity	−1	0	0	−1
Animation	−1	0.5	0.5	−0.5
Structure	1	0	0	1
Context	0	0.5	0.5	0.5
Predicate	−3	1.5	1.5	−1.5

Table 9.2. Scaling and translation factors for the intentional stance

```
;;; Now using a generic function, we represent all the component factors to the disposition to select
;;; this (the physical stance). The factors are implemented by functions which return a value between
;;; zero and one. Before these factors are combined, they are scaled and translated appropriately.

(generic-function (disposition &rest keys)
  (:method ((disposition (eql 'context)) &key &allow-other-keys)
    0.5)
  (:method ((disposition (eql 'predicate)) &key predicate &allow-other-keys)
    (transform (get-physical-predicate predicate)
      :scale -3.0 :translation 1.5)))
```

Figure 9.5. Dispositions for the physical stance

```
;;; The same principles apply, with different scale and translation factors, for the intentional stance.

(generic-function (disposition &rest keys)
  (:method ((disposition (eql 'similarity)) &key from to &allow-other-keys)
    (transform (get-similarity (object-form from) (object-form to))
      :scale -2.0 :translation 1.0))
  (:method ((disposition (eql 'familiarity)) &key from to &allow-other-keys)
    (transform (get-familiarity (object-name from) 'intentional
                 (object-name to) (object-form to))
      :scale -1.0 :translation 0.0))
  (:method ((disposition (eql 'animation)) &key to &allow-other-keys)
    (transform (get-animation (object-form to))
      :scale -1.0 :translation 0.5))
  (:method ((disposition (eql 'structure)) &key from to &allow-other-keys)
    (transform (get-structure (object-name from) (object-name to))
      :scale 1.0 :translation 0.0))
  (:method ((disposition (eql 'context)) &key &allow-other-keys)
    0.5)
  (:method ((disposition (eql 'predicate)) &key predicate &allow-other-keys)
    (transform (get-intentional-predicate predicate)
      :scale -3.0 :translation 1.5)))
```

Figure 9.6. Dispositions for the intentional stance

For the physical stance there are only two dispositions in the model: the context disposition which has a constant energy, and the disposition for the predicate. This context energy corresponds most closely to the social context, in that if it is high then there is a disposition to take some other stance, but if it is low then there is a disposition towards taking the physical stance.

Unlike the physical stance, there are quite a few dispositions that affect the selection of the intentional stance. For the model of Baron-Cohen *et al.*'s false belief test, however, we can disregard, at least for now, some of the more complex aspects, such as the learning processes involved in the development of familiarity. The dispositions that are relevant to this example are the context disposition that models the social context, just like in the physical stance, and the similarity effect. The most important of these is the similarity effect, which promotes selection of the intentional stance for those objects which resemble the person taking the stance. In this example, this means that the intentional stance will generally be taken to humans and to puppets and not to marbles and boxes, but there is no hard barrier to this charmed circle. Even though only one stance can be selected at a time there is still a continuum of similarity.

The physical stance

The physical stance provides a person's ability to make physical predictions. In the context of this model, the physical stance is closely similar to the physical environment model, in that, once again, it is principally concerned with reasoning about the containment of objects one within another. The main difference from the physical environment model is that in the physical stance the rules are more hypothetical—that is, we are concerned with 'what if' reasoning about the results of an action, rather like the possible worlds discussed in chapter 4. This prohibits the direct use of *assert* and *retract* clauses to maintain a database of object containment as used in the physical environment model. This is where the situation calculus comes in; we can look at the effects of an action on a situation without being committed to stay within that new situation.

Another difference between the basic physical model and the kinds of reasoning in the physical stance is the added *place* event handler. A *place* event is generated when a person enters a room, for example, and sees things in particular places. An entire layout can be set up within a room, and

as soon as someone enters that room, they will be able to see each object in its current place. This only applies at the level of a container, though, so if a marble is hidden inside a box when a person enters the room, that person will immediately become aware of the location of the box, but even the existence of the marble will be a secret to her. The behaviour of the *place* predicate with respect to a notional world is rather like that of the *put-in* predicate, in that it may change the believed location of an object within that notional world. The model of a person's physical stance is shown in figure 9.7.

The main part of the physical stance, though, is made up of the rules which are used to make the predictions. These rules are implemented by the *result* function in the situation calculus, which takes the object the stance has been taken to, the action whose effects are to be predicted, and a situation, and returns a new situation describing the effects which the action would bring about on that situation. In this model of the physical stance there are three case rules, once each for the actions of seeing something in a particular place, putting something into a container, and taking it

```
;;; If we see ?object in a place ?container, then we find out where it was in the situation, and return a
;;; new situation so that it is now in ?container and no longer where it was previously, in ?outer-container.

((result yes ?stance-to (place ?object ?container) ?situation ^new-situation) :-
 (member (inside ?object ?outer-container) ?situation)
 (difference ?situation ((inside ?object ?outer-container)) ?situation1)
 (append ?situation1 ((inside ?object ?container)) ?new-situation))


;;; If we see an object being put into a new place, ?container, then again we find out where it was
;;; before in the situation, and return a new situation so that it is now in ?container and no longer
;;; where it was previously, in ?outer-container.

((result yes ?stance-to (put-in ?object ?container) ?situation ^new-situation) :-
 (member (inside ?object ?outer-container) ?situation)
 (difference ?situation ((inside ?object ?outer-container)) ?situation1)
 (append ?situation1 ((inside ?object ?container)) ?new-situation))


;;; If we see an object being taken out of a place ?container, we return a new situation so that it is no
;;; longer in ?container, but is now outside it, in ?outer-container.

((result yes ?stance-to (take-out ?object ?container) ?situation ^new-situation) :-
 (member (inside ?container ?outer-container) ?situation)
 (difference ?situation ((inside ?object ?container)) ?situation1)
 (append ?situation1 ((inside ?object ?outer-container)) ?new-situation))
```

Figure 9.7. The basic physical stance

out of a container. The *member* tests are used to decide whether a given fact is true in the physical description of a situation; the *difference* and *append* primitives are then used to add and remove facts from this original situation, and to construct a new situation with the appropriate changes.

The intentional stance

The intentional stance represents a person's ability to make predictions about the mental states of others. The structure of the intentional stance is broadly similar to that of the physical stance, in that once again it is divided into a set of dispositions that influence the selection of the stance and a set of rules that are used to make predictions once the intentional stance has been selected.

In the model of the intentional stance, there are three main *result* rules. The first rule is associated with the effects of the *perceived* action—this is where the real essence of the intentional stance is represented. This first action does most of the work of modelling Leslie's theory of common-sense psychology, shown in figure 9.1. The other two rules are associated with the effects of the *believes* action, which is only used for asking questions. These rules use the *write-list* primitive to print out the answer to the question, as shown in the traces in Appendix A. The connection with Leslie's model is not completely perfect, though, because Leslie's concern is mainly pretence; some changes to the model are needed to link Leslie's interpreter to the system for physical reasoning.

The first *result* rule in figure 9.8 uses the *ascribe* helper rule to keep all the notional worlds up to date with the predicted effects of the *perceived* event. This *ascribe* rule implements the various elements of the decoupler in figure 9.1, while the central cognitive systems and perceptual processes are represented by the rules belonging to the person as a whole.

The first step in the *ascribe* rule is to use the *those* action to get all the beliefs out of the situation; this set of beliefs corresponds to the ascriber's notional world, and it is this that is passed to the theory of mind mechanism. Next the *requote* action is used; this corresponds to the expression raiser in figure 9.1. In Leslie's examples, the expression raiser can get away with quoting a single statement, but when dealing with more complicated problems than this, such as Baron-Cohen *et al.*'s (1985) false belief test, Dennett's concept of a notional world comes back into its own. In this

model, then, the role of the expression raiser is actually played by a notional world raiser, that selects and raises a notional world as a unit, rather than dealing with a single statement. The *requote* action takes one list describing a notional world and returns a second, so that it is the whole notional world that is raised, before being passed to the manipulator. During the raising process in *requote*, each of the first state's elements is matched against a pattern, and its bindings are used to instantiate a second pattern to construct the corresponding element in the raised notional world.

Finally, the manipulator is implemented by the *in-stance* relation; this step selects the appropriate stance to ascribe the event to someone or something. Once again, *in-stance* is a special primitive in the modelling language, taking an object, a prediction, and a query. It operates by taking the appropriate stance to the given object for the given prediction, and then evaluating the given query in that context. This last step, evaluating the query, corresponds to passing the query on to

```
;;; When we perceive something and are taking the intentional stance to ?stance-to, get ?stance-to's
;;; notional world into ?notional-world, and then, in that model, take the right stance for the event,
;;; predicting its physical effects. Then map these effects into changes to ?stance-to's notional world.

;;; Rule perceive
((result ^response ?stance-to
         (perceived ?object (?other-action ?other-object ?event))
         ?situation ^new-situation) :-
 (ascribe ?stance-to ^response ?stance-to
          (perceived ?object (?other-action ?other-object ?event))
          ?situation ^new-situation))

;;; Rule ascribe
((ascribe ?someone ^response ?stance-to
          (perceived ?object (?other-action ?other-object ?event))
          ?situation ^new-situation) :-
 (those (believes ?someone ?something) ?situation ?notional-world)
 (requote (believes ?someone ?something) ?notional-world ?something ?situation2)
 (in-stance ?other-object ?other-action
   (result ?response ?other-object (?other-action ?other-object ?event)
           ?situation2 ?situation3))
 (requote ?something ?situation3 (believes ?someone ?something) ?new-notional-world)
 (difference ?situation ?notional-world ?situation1)
 (append ?new-notional-world ?situation1 ?new-situation))

;;; These are the rules for answering questions about people's beliefs. Do this by looking for the truth
;;; of the proposition in ?object's notional world.

;;; Rule answer-yes
((result yes ?stance-to (believes ?object ?something) ?situation ^situation) :-
 (member (believes ?object ?something) ?situation)
 (write-list (yes ?object believes ?something)))

;;; Rule answer-no
((result no ?stance-to (believes ?object ?something) ?situation ^situation) :-
 (not (member (believes ?object ?something) ?situation))
 (write-list (no ?object does not believe ?something)))
```

Figure 9.8. The basic intentional stance

the interpreter in figure 9.1. In the context of this selected stance, the interpreter is used as a nested *result* action. This returns a new situation, which is then again passed to the *requote* action, to restore its quotation status to normal as a notional world. Finally, this new notional world is used to replace the one in the original situation, and the changed situation is returned.

The other two *result* rules in figure 9.8, for the *believes* query, model those parts of the cognitive model which answer questions. They are much simpler, because they never need to change a situation; they only need to look for the presence or absence of a belief in the notional world. These rules both return the same situation that they are passed, but they also return either *yes* or *no* as the response, depending on whether or not the ascriber believes the statement passed in.

Perhaps this will be clearer with a more concrete example. Imagine that the given query is *(result ?response sally (perceived sally (put-in marble box)) ?situation ?new-situation)* in a situation S. Because the object to take the stance to is *sally*, and the prediction is *perceived*, this will almost certainly result in taking the intentional stance and the appropriate *result* rule will be selected and applied, calling the *ascribe* rule. The *those* action and the first *requote* action are then used to go through the situation S, decoupling all the relations *(believes sally ?something)*, and generating a new situation S' containing the entire contents of Sally's notional world. This done, the model takes the appropriate stance to the *marble* for the prediction *(put-in marble box)*, but now in the situation S'—and this time it will be the physical stance that is chosen. The physical stance and physical prediction rules will then be used to generate a new situation R'. The second *requote* action is then used to go through R' and restore its quotation status to normal, in a new situation R. Finally, the new situation, R, is used to replace the original situation, S, and is returned.

The connection between the intentional stance and the physical stance here is provided by the special relation *in-stance*, which effectively nests one prediction inside another. This implementation raises some big issues which are only sketchily addressed by the psychological studies which underpin this theory. In Leslie's model, the only place where physical reasoning takes place is the manipulator—but one of the curious effects here is that the manipulator can here end up with a psychological prediction rather than a physical one. It could be, for example, that the query posed was a nested psychological prediction like *(believes sally (believes anne (inside marble ?where)))*. These

kinds of nested queries and problems require several levels of coupling and decoupling, which are not clearly accounted for in the apparently modular descriptions favoured by Leslie. Nevertheless, even in theories such as Leslie's (1995) and in Baron-Cohen *et al.*'s (1985) false belief test, intentional and physical reasoning are both required and must be properly linked. The use of nested stances in this model will let this happen, but the approach needs a lot more investigation; it may well be that far more subtle and complex systems are involved in questions like this. This issue also shows up in more philosophical analyses of common-sense psychology, where it is commonly claimed (e.g. Dennett, 1987) that a distinguishing feature between people and animals is that people are *n*th order intentional systems, where *n* is greater than about three or four. Dennett admits that people are, in practice, limited to depths of nesting about five or six deep. But Dennett's examples show an increasingly 'game-like' tendency that seems to imply that although people naturally handle the lower orders of intentionality, quite different mechanisms may well be involved in these higher orders—mechanisms that only humans may possess. The proper character of people's reasoning about higher order intentional states, and the connection between the intentional stance and the physical stance, still need to be clarified. This model may not have the answer, but at least it has raised the question.

Summary

In this chapter I have described the basic model of common-sense psychology. This includes elements which correspond to a basic common-sense physics as well as a basic common-sense psychology, along with a simple model of anthropomorphism—of the dispositional factors which influence the selection of one stance over another. This model corresponds roughly to Leslie's model of the 'theory of mind mechanism', although there are a number of significant differences which are needed to accommodate the combination of physical and psychological reasoning that makes proper handling of mental ascription possible.

In the rest of this part of the thesis, I'll show how this model can be used; first, to study a child's ascription of beliefs in Baron-Cohen *et al.*'s false belief test, and second, to study Searle's general ascription of mentality in the thought experiment that accompanies his Chinese Room argument.

Chapter 10
Modelling the false belief test

Introduction

I have already described the theory behind Baron-Cohen *et al*'s (1985) false belief test in chapter 3, and in the previous chapter, I showed how we can use the theories I have outlined to model people's common-sense psychology through taking the intentional stance. In this chapter, I will take these ideas further, and connect them to develop a relatively complete computational model of the processes involved in the false belief test. Once again, I should stress that this model is intended to be descriptive rather than an accurate model of the actual processes that are involved.

At this point, we have a model for a complete, though rather simple, notional world theory of mind, corresponding roughly to Leslie's (1995) 'theory of mind mechanism' described in the previous chapter. This has been selected as much for its clarity as for its empirical correctness—I'll present more empirically correct variations on the model later in the chapter, and compare them with this first version. In general, these will all be represented as minor variations in the rules for the intentional stance. The models I'll use for comparison are versions of the simulation theory (Goldman, 1993; Gordon, 1986), of Chandler's (Chandler & Boyes, 1982) 'copy theory', and of Perner's (1991) 'situation theory'. From the differences and similarities between these models, I'll show that we can draw some conclusions about the corresponding differences between the theories they represent, and therefore about the strengths and weaknesses of these theories, and of the modelling approach.

Modelling Baron-Cohen *et al.*'s false belief test

First of all, I will describe the different parts of the model of Baron-Cohen *et al.*'s false belief test in more detail, and show how they act together to build up a picture of the child's theory of mind. There are three parts to the complete model of the false belief test, first we need to describe and introduce all the objects involved to the modelling environment, then we need to run a script for the actions in the false belief test, and finally we need to ask some questions of the model, to tell us whether or not our modelled subject passed the test.

In a model of the false belief test, the objects that we need to describe correspond to the puppets Sally and Anne, the marble, the basket and the box, and the child subject, who we can call Alison. The descriptions of these objects are shown in figure 10.1. Each of the objects defined in figure 10.1 is modelled by giving them a name and a corresponding physical form, and optionally, an object database and a set of stances. The physical form names are only used to define the physical feature attributes for these objects, so the model can use the similarity measures discussed in chapter 8; these forms and their corresponding feature attributes are described fully in figure 9.3, and defined in the file *forms.component* in Appendix C. The symbolic name of the form is not used anywhere to label the class of object involved—there is no class *doll*, for example, in the sense of object-oriented programming.

```
;;; Declare the objects, along with their appropriate physical forms, for the false belief test.

(model-object sally :form doll)
(model-object anne :form doll)
(model-object marble :form marble)
(model-object basket :form basket)
(model-object box :form box)


;;; Alison is modelled similarly, but with additional features.  Because Alison is a person, we not only
;;; specify that form, we also give her a database with perceptual apparatus, and stances for basic
;;; physical and intentional reasoning.

(model-object alison :form person
  :stances (basic-intentional-stance
            basic-physical-stance)
  :database basic-person-object-database)
```

Figure 10.1. Object declarations for the false belief test

The only active participant in the model, then, is the child subject, Alison. The object description for Alison is different because she is given the appropriate object database and stances for a person, giving her the perceptual systems and reasoning processes appropriate for a person with an adult theory of mind, described by the model in the previous chapter and corresponding to Leslie's (1995) theory of mind mechanism model.

When the objects needed are properly known to the model, we can begin to act out Baron-Cohen *et al.*'s false belief test. We do this by telling the model about the physical actions that correspond to the movements of objects in the world. The *tell-model* primitive is used for this. It passes the action to the physical environment model, which then works out the perceptions that should arise from this action. Every object is notified of its perceptions, and Alison, in particular, will use this to update her notional worlds for all the objects that she knows about.

There are two parts to the physical actions in the false belief test, shown in figure 10.2. The first part consists of a set of actions which simply introduces the various objects to the model, one at a time, in any order. Finally, Alison is put into the room. As soon as she enters the room, she becomes aware of all the other objects that are already there, and will, when appropriate, ascribe them notional worlds made up of beliefs about the places of these objects.

As soon as this has been done, the main part of the false belief test can be carried out. We begin by putting the marble into the basket, with Sally and Anne both present, and Alison watching. In Alison's notional worlds, then, both Sally and Anne will come to believe that the marble is in the basket. Sally then leaves the room. While Sally is outside the room, the marble is moved to the box, by first taking it out of the basket and then putting it into the box, again with Alison watching. Then Sally comes back into the room. This is the critical point in the false belief test, because now Alison's notional worlds for Sally and Anne should be different. In Sally's notional world, the marble should still be in the basket, but in Anne's (and Alison's) notional world the marble should now be in the box. We can then use the model primitive *ask-object-if* to ask Alison the questions that Baron-Cohen *et al.* used to look at these notional worlds. These questions are shown in figure 10.3.

The answers to these questions tell us how the child subject in this particular model responded to the false belief test. The answer to the second question is the distinguishing one; a child is capable of ascribing false beliefs to others if the answer to the second question is different from the answers to the first and the third questions. In the first instance, with Leslie's theory of mind mechanism in place, Alison will pass the false belief test by giving different (and appropriate) answers to the second and third questions.

Note that the difference between Sally's and Anne's notional worlds is not simply due to Sally missing out on changes to her database because she is out of the room. The changes Sally misses out on are changes to her notional world, which is actually part of Alison's database, and Alison

```
;;; We start by introducing the characters here. The order doesn't matter much. Alison will become
;;; aware of all the other objects as soon as she enters the room.

(tell-model (put-in basket room))
(tell-model (put-in box room))
(tell-model (put-in marble room))

(tell-model (put-in sally room))
(tell-model (put-in anne room))

(tell-model (put-in alison room))

;;; Now the scenario starts. Put the marble in the basket
(tell-model (put-in marble basket))

;;; Sally leaves the room
(tell-model (take-out sally room))

;;; Move the marble from the basket into the box
(tell-model (take-out marble basket))
(tell-model (put-in marble box))

;;; Sally comes back into the room
(tell-model (put-in sally room))
```

Figure 10.2. Actions for the false belief test

```
;;; Where does Alison think that the marble is?
(ask-object-if alison (believes alison (inside marble ?where)))

;;; Where does Alison think that Sally thinks the marble is?
(ask-object-if alison (believes sally (inside marble ?where)))

;;; Where does Alison think that Anne thinks the marble is?
(ask-object-if alison (believes anne (inside marble ?where)))
```

Figure 10.3. Asking Alison where the marble is in the false belief test

was present all along. Neither Sally nor Anne, being dolls, have databases at all. Sally's notional world is spared the changes that happened to Anne's because Alison is aware that she is out of the room while the marble is moved from the basket to the box. Alison only ascribes awareness of events to agents she can see (hence the *sees* relation in the ***handle-event*** rule in figure 9.3) and it is this, with the ***ascribe*** rule in figure 9.8, that leads to the difference between Sally's and Anne's notional worlds.

Comparing models: the simulation theory

The first model I'll compare to the theory of mind mechanism is a version of the simulation theory. The simulation theory is typified by a 'role taking' or 'perspective taking' approach. Gordon describes this strategy as saying that "*Smith believes that* Dewey won the election" should be read as "let's do a Smith simulation. Ready? *Dewey won the election*" (Gordon, 1986).

The model for the simulation theory has a slightly modified intentional stance, shown in figure 10.4. The rules in figure 10.4, then, replace the original ***perceive*** rule for predicting the effects of a perceived event shown in figure 9.8. The idea here is that, for oneself, ascription proceeds as before. Initially, however, the subject is not able to ascribe mental states to others. This is shown by an empty second rule, which does nothing.

```
;;; The rules for the first version of the simulation theory. Initially, if we are seeing something
;;; ourselves, then we do the right ascription, otherwise we leave the situation alone. These two rules,
;;; together, replace the perceive rule in the basic intentional stance shown in figure 9.8.

;;; Rule perceive-self, compare to perceive in figure 9.8
((result ^response ?stance-to
         (perceived ?object (?other-action ?other-object ?event))
         ?situation ^new-situation) :-
 (self ?stance-to)
 (ascribe ?stance-to ?response ?stance-to
   (perceived ?object (?other-action ?other-object ?event))
   ?situation ?new-situation))

;;; Rule perceive-other, compare to perceive in figure 9.8
((result ^response ?stance-to
         (perceived ?object (?other-action ?other-object ?event))
         ?situation ^situation) :-
 (not (self ?stance-to)))
```

Figure 10.4. Rules for the simulation theory

When run (see *mfbsimulationfail.trace* in Appendix A), this seems to fail the false belief test correctly in that Alison simply doesn't give answers at all for either Sally or Anne; before Alison can pass the test she needs to acquire the ability to simulate, or take the role of, other people. This corresponds to the development of a simulation competence: "before internalising this system, the child would simply be unable to predict or explain human action [but] after internalising the system the child could deal indifferently with actions caused by true beliefs and actions caused by false beliefs" (Gordon, 1986). This is why the kind of failure in this simulation model is interesting; Alison simply fails to give answers for either Sally or Anne, because she failed to take their roles properly.

The second stage in the model, then, is the complete simulation rule, which implements a role taking strategy through the *in-self* primitive. This primitive has the effect of temporarily pretending to be a different self, and then handling the whole event in that context instead. It is this replacement second ***perceive*** rule that allows Alison to pass the false belief test. The replacement rule which models this role taking strategy is shown in figure 10.5.

One point that should be clarified here, though, is the relationship between the physical and intentional stances. Because it uses the *in-stance* primitive, in a sense, even the model of Leslie's theory of mind mechanism uses simulation as a connection between intentional and physical reasoning, as do all the other models. The actual form of the connection between physical and intentional reasoning hasn't really been studied in enough detail for clear models to be possible. However, using simulation as a tool for physical reasoning is realistic (Clark, 1989), so it is legitimate for all the different models to use something like simulation for this connection. The key identifying

;;; The replacement second rule for the simulation theory. If we are not seeing something for ;;; ourselves, then we "pretend" to be someone else through the *in-self* primitive, and process the ;;; event as if we were that person. This rule replaces the ***perceive-other*** rule in figure 10.4.

```
;;; Rule perceive-other, compare to perceive-other in figure 10.4.
((result ^response ?stance-to
         (perceived ?object (?other-action ?other-object ?event))
         ?situation ^new-situation) :-
 (not (self ?stance-to))
 (in-self ?stance-to
   (result ?response ?stance-to
     (perceived ?object (?other-action ?other-object ?event))
     ?situation ?new-situation)))
```

Figure 10.5. Replacement second rule for the simulation theory

feature of a simulation model, then, is the recursive use of a *result* rule in a different context—either in a different role for intentional reasoning (using the *in-self* primitive) or in a different stance for physical reasoning or nested intentional reasoning (using the *in-stance* primitive).

There are a number of important conclusions to be drawn from this idea. First, in the simulation model the behaviour involved in ascribing mentality to oneself seems to be different from that involved in ascribing mentality to others. This contrasts with the theory of mind mechanism described in the previous chapter, where there is no difference between first person and third person ascription. This is shown by the rules' sensitivity to the *self* relation. This shows that there is an egocentricity involved in the simulation theory. The second point to note is that, in practice, in adults at least, the behaviour of this system is the same as that of the basic theory of mind mechanism anyway, because the new *perceive-other* rule combines with the *perceive-self* rule to behave just as if there was a single rule which uses the *ascribe* action—a rule which is identical to the first *result* rule in the intentional stance model in figure 9.8. This is in accord with Dennett's (1987) point that, in practice, the difference between a theory and a simulation may be "at worst one of emphasis".

Comparing models: the copy theory

The second model I'll compare against Leslie's theory of mind mechanism is Chandler's 'copy theory'. Chandler and Boyes describe younger children as behaving "as though they believe objects to transmit, in a direct-line-of-sight-fashion, faint copies of themselves which actively assault and impress themselves upon anyone who happens in the path of such 'objective' knowledge" (Chandler & Boyes, 1982). They argue that this is the precursor to a complete theory of mind like Leslie's, and therefore I'll only show the version which fails the false belief test—a version which passed the test would be identical to Leslie's theory of mind mechanism as described in the previous chapter.

From the complete model of the theory of mind mechanism corresponding to an adult theory of mind, then, we can modify the intentional stance slightly to represent a child with a copy theory of belief. The idea behind the copy theory is, in effect, that instead of ascribing beliefs to others, a

'copy' of one's own beliefs is used instead. Instead of building different notional worlds for Sally and Anne, both have notional worlds which are copies of Alison's. There are two parts to the changes for the copy theory; these changes are shown in figures 10.6 and 10.7.

According to the copy theory, children simply do not ascribe real beliefs to others. This is shown by the modified *perceive* rules in figure 10.6, which replace the *perceive* rule in figure 9.8 so that beliefs are only ascribed to oneself. Note that these new rules are identical to those of the simulation model in figure 10.4; this is to be expected—Chandler's theory is an account of how children escape from the kind of egocentricity that marks a simulation theory. But this is not the whole

```
;;; The ascription rules for the copy theory.  Initially, if we are seeing something ourselves, then we
;;; do the right ascription, otherwise we leave the situation alone.  These two rules, together, replace
;;; the perceive rule in the basic intentional stance shown in figure 9.8.  Note that these replacement
;;; rules are identical to those in figure 10.4.

;;; Rule perceive-self, compare to perceive in figure 9.8
((result ^response ?stance-to
         (perceived ?object (?other-action ?other-object ?event))
         ?situation ^new-situation) :-
 (self ?stance-to)
 (ascribe ?stance-to ?response ?stance-to
   (perceived ?object (?other-action ?other-object ?event))
   ?situation ?new-situation))

;;; Rule perceive-other, compare to perceive in figure 9.8
((result ^response ?stance-to
         (perceived ?object (?other-action ?other-object ?event))
         ?situation ^situation) :-
 (not (self ?stance-to)))
```

Figure 10.6. Object ascription rules for the copy theory

```
;;; The answer rules for the copy theory.  These have the effect of considering the target's notional
;;; world to be a 'copy' of the ascriber's.  These rules replace the rules answer-yes and answer-no in
;;; the basic intentional stance shown in figure 9.8.

;;; Rule answer-yes-self, compare to answer-yes in figure 9.8
((result yes ?stance-to (believes ?object ?something) ?situation ^situation) :-
 (self ?self)
 (member (believes ?self ?something) ?situation)
 (write-list (yes ?object believes ?something)))

;;; Rule answer-no-self, compare to answer-no in figure 9.8
((result no ?stance-to (believes ?object ?something) ?situation ^situation) :-
 (self ?self)
 (not (member (believes ?self ?something) ?situation))
 (write-list (no ?object does not believe ?something)))
```

Figure 10.7. Answer rules for the copy theory

story in the copy theory; when children are asked about other people's beliefs, they answer by drawing on their own beliefs. For this, we also need to change the *answer* rules in figure 9.8; these are the rules which model how the child answers the kinds of question used in the false belief test. These changes are shown in figure 10.7. Both the *answer* rules are changed from figure 9.8 by using the *self* relation to find and use one's own beliefs, rather than anybody else's, to answer the given question. This dependence on the *self* relation is important—again it shows that there is an egocentricity in the copy theory, just as there is in the simulation theory.

There are more complex variations on the 'copy theory', for instance Wellman (1990) argues that younger children have a copy theory of belief, but not of desires. This variation remains outside the scope of this model because desire psychology isn't yet part of the modelling environment—this is an area for future work. But while the copy theory model works to the extent that, when run, it correctly fails the false belief test, the model is quite radically different from that of an adult theory of mind, and it does seem to require a developmental jump of significant magnitude. All the egocentricity of the rules in figure 10.4 and 10.5 must be lost, and the child needs to learn to extend notional worlds to other people. This matches all the empirical evidence that is against a copy theory; Perner (1991) has argued quite convincingly that experiments involving inference from parts to wholes show that the evidence is against children having a copy theory at any age. This is something which could, in principle, be investigated further quite easily with this modelling approach.

Comparing models: the situation theory

The third reference comparison I'll make against the theory of mind mechanism is Perner's 'situation theory'. Perner's theory is substantially different from those presented so far, because he draws a hard distinction between real and non-real situations, or contexts. The notional world an agent has of itself has a rather unique status. This is not mirrored in the basic model of the intentional stance in figure 9.8, where all the notional worlds have the same status.

Perner (1991) argues that the reason younger children don't pass the false belief test is because the child subject applies the verbal form of questions incorrectly to the situation corresponding to reality, not to the non-real situation which has been acted out by the puppets. According to the situation theory, unlike the copy theory, young children do have notional worlds, but they are not so good at understanding that a real question can apply to a non-real situation. Perner uses this distinction to explain why children who can't pass the false belief test are still capable of sophisticated notional world reasoning, such as that required by Zaitchik's (1990) 'false photograph' test.

Figures 10.8 and 10.9 show the rules for the first version of the situation theory model. Note that the *perceive* rule has now been split into two: one for self and one for others—creating notional worlds with different predicates, *knows* for oneself, and *believes* for others. These represent the

```
;;; The ascription rules for the situation theory.  These add a status flag to the rules which ascribe
;;; notional worlds.  This status value is knows for one's own notional world, and believes for other
;;; people's.  These two rules, together, replace the perceive rule in the basic intentional stance shown
;;; in figure 9.8.

;;; Rule perceive-self, compare to perceive in figure 9.8
((result ^response ?stance-to (perceived ?stance-to (?other-action ?other-object ?event))
         ?situation ^new-situation) :-
 (self ?stance-to)
 (ascribe ?stance-to knows ?response ?stance-to
   (perceived ?stance-to (?other-action ?other-object ?event))
   ?situation ?new-situation))

;;; Rule perceive-other, compare to perceive in figure 9.8
((result ^response ?stance-to
   (perceived ?object (?other-action ?other-object ?event))
   ?situation ^new-situation) :-
 (not (self ?stance-to))
 (ascribe ?stance-to believes ?response ?stance-to
   (perceived ?stance-to (?other-action ?other-object ?event))
   ?situation ?new-situation))

;;; The ascription rule is extended to take the additional status value.  This value is used, instead of
;;; the fixed status value believes, to distinguish between one's own notional worlds and other people's.
;;; This rule replaces the ascribe rule in the basic intentional stance shown in figure 9.8.

;;; Rule ascribe, compare to ascribe in figure 9.8
((ascribe ?someone ?status ^response ?stance-to
   (perceived ?object (?other-action ?other-object ?event))
   ?situation ^new-situation) :-
 (those (?status ?someone ?something) ?situation ?notional-world)
 (requote (?status ?someone ?something) ?notional-world ?something
   ?situation2)
 (in-stance ?other-object ?other-action
   (result ?response ?other-object
     (?other-action ?other-object ?event) ?situation2 ?situation3))
 (requote ?something ?situation3 (?status ?someone ?something) ?new-notional-world)
 (difference ?situation ?notional-world ?situation1)
 (append ?new-notional-world ?situation1 ?new-situation))
```

Figure 10.8. Ascription rules for the situation theory

different statuses for the corresponding notional worlds that is essential to Perner's theory. Initially, as presented in figure 10.9, children can only link verbal questions to the world for self beliefs. Other notional worlds can and do exist, though; it is just that they cannot be accessed through verbal questions. Note that there is something like egocentricity even in Perner's account—the rules in figures 10.8 and 10.9 are sensitive to the *self* relation, just as in the simulation theory and the copy theory. The difference is that there is no qualitative difference between the processes for first person and third person ascription; the only difference is in the status of the notional worlds that they generate.

Perner claims that the principal change in children between the ages of two and a half and four is the acquisition of a representation theory, which allows them to recognise that questions like these can refer not to reality, but to worlds or situations that are represented; that is, the change is from the child as a situation theorist to the child as a representation theorist. I'll show this development in a slight modification of the ***answer*** rules in figure 10.9, to the rules in figure 10.10.

Perner argues that this change isn't a radical overturning of the existing theory. Instead, he suggests that the change that happens at around this time is a "theory extension" (Perner, 1991), a relatively minor change to the existing theory. This accords well with the interpretation of the word 'theory' that I have adopted in this thesis—not as something scientific that is either right or wrong in a black-and-white sense, but as something rather more elastic, that is capable of refinement and adjustment within a loose framework set by its predictive utility. This character of theory extension should be an important part of any developmental account of common-sense psychology, because the empirical evidence is that common-sense psychology develops gradually, not in big jumps (Carey, 1985).

Discussion

These models highlight several of the most important features of the common-sense psychology that underlies the false belief test, and show that these features can be investigated by models that represent the different and competing theories in this field. Of the models presented, the one that seems to work best in this modelling framework is Perner's 'situation theory'. The principal rea-

```
;;; Situation theory rules for answering questions about one's own beliefs. In this group, the "believes"
;;; question is coupled to the knows predicate of notional world.  These implement the 'self' half of
;;; the answer rules in the basic intentional stance shown in figure 9.8.

;;; Rule answer-yes-self, compare to answer-yes in figure 9.8
((result yes ?stance-to (believes ?self ?something) ?situation ^situation) :-
 (self ?self)
 (member (knows ?self ?something) ?situation)
 (write-list (yes ?self believes ?something)))

;;; Rule answer-no-self, compare to answer-no in figure 9.8
((result no ?stance-to (believes ?self ?something) ?situation ^situation) :-
 (self ?self)
 (not (member (knows ?self ?something) ?situation))
 (write-list (no ?self does not believe ?something)))

;;; Situation theory rules for answering questions about other people's beliefs. This is a model of what
;;; happens before the representation theory is acquired, where the effect is to link into the knows
;;; predicate instead of the believes predicate.  These implement the 'other' half of the answer rules
;;; in the basic intentional stance shown in figure 9.8.

;;; Rule answer-yes-other, compare to answer-yes in figure 9.8
((result yes ?stance-to (believes ?object ?something) ?situation ^situation) :-
 (not (self ?object))
 (member (knows ?self ?something) ?situation)
 (write-list (yes ?object believes ?something)))

;;; Rule answer-no-other, compare to answer-no in figure 9.8
((result no ?stance-to (believes ?object ?something) ?situation ^situation) :-
 (not (self ?object))
 (not (member (knows ?self ?something) ?situation))
 (write-list (no ?object does not believe ?something)))
```

Figure 10.9. Answer rules for the situation theory

```
;;; Situation theory replacement rules for answering questions about other people's beliefs.  In this
;;; group, the "believes" question is correctly coupled to the believes predicate of notional world.
;;; These rules override the default which gives the wrong answer in the first version of the situation
;;; theory.

;;; Rule answer-yes-other, compare to answer-yes-other in figure 10.9.
((result yes ?stance-to (believes ?object ?something) ?situation ^situation) :-
 (not (self ?object))
 (member (believes ?object ?something) ?situation)
 (write-list (yes ?object believes ?something)))

;;; Rule answer-no-other, compare to answer-no-other in figure 10.9.
((result no ?stance-to (believes ?object ?something) ?situation ^situation) :-
 (not (self ?object))
 (not (member (believes ?object ?something) ?situation))
 (write-list (no ?object does not believe ?something)))
```

Figure 10.10. Changes from the situation theory to the representation theory

son for this is that the apparent distance between passing and failing the false belief test is much smaller. For both the simulation theory and for the copy theory there must be a radical development to the ascription of notional worlds, and initially there must be a slightly different intentional stance which ascribes one's own beliefs to others. Perner's model shows best the character of "theory extension" (Perner, 1991), a character which, he suggests, should be expected of a theory which matches the empirical psychological data of the development of these theories.

The simulation theory is actually quite similar to the version of Leslie's theory of mind mechanism that we have used as a base model—but both it and Chandler's copy theory show an apparent egocentricity. In practice, as I've already argued, there are good reasons for supposing that in any real common-sense psychology, both theory and simulation aspects will be required, and, therefore, a simulation theory should really be complementary to, rather than alternative to the models presented here (Perner, 1994). However, many of the people who have argued for a simulation theory have argued for it as an alternative to a theory of mind, and therefore don't leave much thought to how a simulation theory and a theory theory might be combined in practice. But there is an important twist to the simulation model in this thesis; although it shows an apparent egocentricity, it is actually functionally identical to the model of Leslie's theory of mind mechanism. This seems further to back up the arguments that the distinction between a theory and a simulation is one of interpretation rather than a real difference in behaviour.

It is, of course, possible to pursue this strategy still further, developing models of some of the other models of theory of mind. Unfortunately, for an accurate model many of these require more complex models of perceptual apparatus (e.g. Baron-Cohen's, 1995, shared attention mechanism), or richer models of common-sense psychology (e.g. Wellman's, 1990, simple desire psychology) than have yet been developed within this framework. Even so, as a first look at the problem, the technique does seem to back up the existing points and arguments remarkably well, and to clarify the distinctions between the models which have been developed so far. And apart from anything else, at least within this limited scenario, it seems to work!

The usefulness of the modelling approach as a tool for studying common-sense psychology is a topic which deserves fuller discussion, and one to which I will return in the last part of the thesis, and particularly in chapter 14. But before this usefulness can be properly assessed, we should look in a bit more detail at some other aspects of the theoretical model developed in the previous part of the thesis; aspects which are not fully revealed in the false belief test. To look at these in more detail, I will return to what seems like a very different problem—the Turing test—and show how it too is a vehicle for common-sense intuitions. I'll do this by looking at one exceptionally rich thought experiment on the theme of the Turing test, Searle's Chinese Room.

Chapter 11
Intuition in the Chinese Room

Introduction

The false belief test evaluates one very specific kind of ascription of mental states. In this chapter, and the one that follows, I'll broaden this out to investigate a much more general kind of ascription—the ascription of mentality as a whole. This brings us back to the Turing test discussed in chapter 5, which is one of the most thoroughly investigated frameworks for deciding whether or not something has a mind, at least in the field of cognitive science. Adopting the Turing test as a paradigm instance of human computer interaction has some interesting repercussions for cognitive science in general; I'll come back to these in chapters 13 and 14. As the review of the Turing test in chapter 5 showed, interpreting the test inductively rather than operationally makes a lot more sense, and allows us to use the format of the test to gather information about the behaviour of a system.

In this chapter and the next I will expand on one specific variation on the Turing test—Searle's (1980) Chinese Room argument. It is not my intention to turn this into a long philosophical debate—in the years since Searle's argument was first published the debate has raged pretty well constantly as it is. Instead, I want to focus on one particular aspect of Searle's article, namely the thought experiment at its heart. This implies a strong distinction between Searle's philosophical argument and the thought experiment which underpins it—a distinction that makes sense given the clear split in the replies between those who attack Searle's philosophical argument on the nature of semantics and those who attack his use of intuitions in the thought experiment underneath. Of course, the role of intuition in philosophical discourse may be an important one, but for safety—even in philosophy—a proper account of the effects of these intuitions is required, and this account simply seems to be missing.

I will not go into any detail regarding Searle's philosophical argument, which is, I think, not fundamentally sound; for example, it does seem to be vulnerable to slippery slope arguments in the manner of Dennett, as described in chapter 2. But much of this argument depends on whether or not you accept the idea of original intentionality being categorically different from derived intentionality—and since this philosophical debate isn't really deeply connected with intuition, it is only of indirect interest in this study of the Chinese Room.

Analysing philosophical thought experiments like this might seem to be something of a digression from the rest of the thesis, but that is not the case. I have already discussed the importance of philosophical discourse to common-sense psychology, but in this chapter I will close the loop, and show that conversely, common-sense psychology is of fundamental importance to philosophical discourse. The Chinese Room thought experiment is designed to appeal to our intuitions, not to argument, and these intuitions are one manifestation of our common-sense reasoning—our judgement and interpretation of Searle's thought experiment are coloured by our human common-sense psychology. In fact, I will argue later (in chapter 14) that many philosophical thought experiments, psychological models, and metaphors are interpreted through the prejudices of our anthropocentric perceptions, and need to be treated with caution for this reason.

But more than this, I am looking at Searle's thought experiment because it throws a new light on common-sense psychology, and clarifies some of the factors involved. This is because it offers a clear, concise, and consistent description of a system with which we are asked to identify. The factors and principles involved in this identification are remarkably revealing. It is for these reasons that Searle's thought experiment forms the second computational model in this thesis. This chapter will lay down some of the issues that are involved and fill in some of the psychological gaps left in the predominantly philosophical analyses of Searle's article. The model itself will be described more fully in the next chapter.

An overview of the Chinese Room

Searle's Chinese Room argument was originally intended to provide an *a priori* argument against the possibility of a running computer program being sufficient to understand natural language, but since then a second thread has been added to the argument, namely that it is impossible for a running computer program to be sufficient for consciousness. The importance of this second thread will become clearer, as it is here that Searle particularly insists that we take the first person stance to the system. This is an essential element of the intuitive appeal of his thought experiment.

The thought experiment is really a variation on the Turing test, but with frills. The first frill is that the dialogue between the observer and the system is in Chinese rather than in English: a language Searle doesn't understand. For the system, Searle imagines a room, where questions can be put through a kind of letter box and answers received in the same way. I've presented an external view of the people outside interacting with the room in figure 11.1a.

The main difference from a standard Turing test is that Searle tells us what is actually going on inside the room—and here his description is intended to closely parallel Schank's (e.g. Schank & Abelson, 1977) model of language understanding. Inside the room is an agent, Searle-in-the-room, along with a whole set of slips of paper with Chinese symbols on them. Searle-in-the-room also has a rule book, written in English, to tell him how to use these slips of paper; so that, for example,

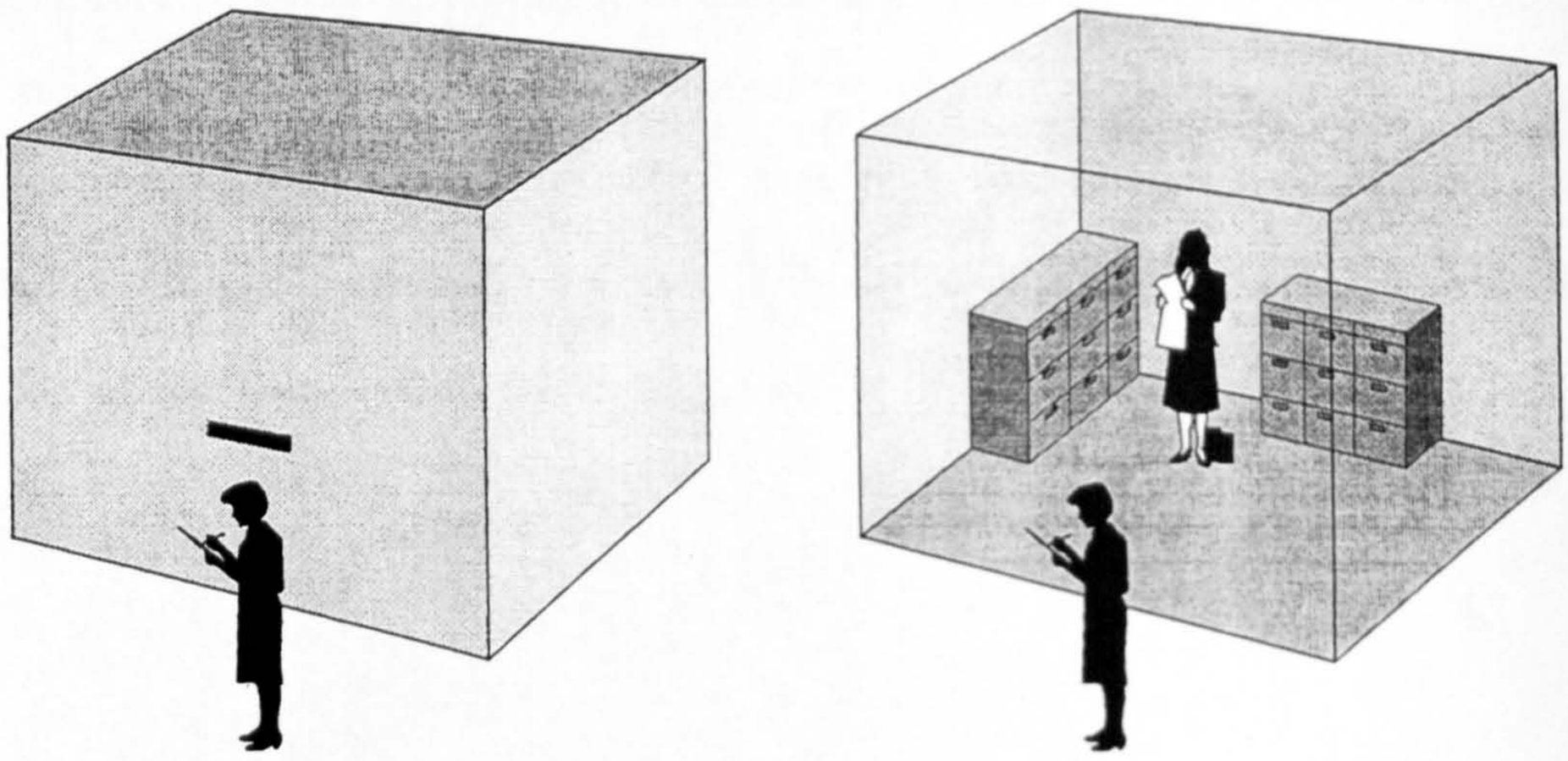

Figure 11.1. The Chinese Room; (a), the external view, and (b), the internal view

if 'squiggle squiggle' is passed in to the room, he should pass out 'squoggle squoggle', although the rules will usually involve much longer chains of pattern matching before anything is passed out of the room. The idea is that Searle-in-the-room can use this rule book and all the slips of paper to generate the appearance of understanding Chinese sufficiently well that the observer outside the room believes that the room really does understand Chinese. This is why the dialogue is conducted in Chinese: it provides a clear separation between the kind of understanding Searle-in-the-room has (which is of English), and the kind of understanding the room shows externally (which is of Chinese). I've presented a view of what might be going on inside the room in figure 11.1b.

Searle's point is that Searle-in-the-room doesn't understand Chinese—so he asks where the real understanding is? In fact, Searle argues that there is, in truth, no understanding there at all, and that no matter how well the room seems to understand Chinese, in practice it never really does. The underlying philosophical argument is that because Searle-in-the-room, through the rules, can only manipulate the Chinese symbols syntactically and not semantically, there are no semantic processes (and, therefore, there is no understanding) in the behaviour of the room.

In his original article (Searle, 1980), Searle describes and rebuts a number of the more obvious replies to this argument I briefly discussed some of these in chapter 5, namely the "systems reply", the "robot reply", and the "other minds reply". All of these appeal to common-sense psychological intuitions in different ways, so they are central to an understanding of the role of intuition in Searle's argument. There are other replies to Searle's argument, both in his original paper and its accompanying commentary, and more have been written since, but many of these are philosophical in nature and are aimed principally at the philosophical argument, not at the intuitive aspects of the thought experiment underneath. Because they do not illuminate people's common-sense intuitions in useful ways, I will not address them further here. But of the initial replies, perhaps the other minds reply is most interesting to this thesis because, as I have already suggested in chapter 5, there is a close connection between the philosophical 'other minds' problem that underpins this reply, and the psychological other minds problem that common-sense psychology sets out to deal with.

The main point I want to draw from all of these replies is that even though the structural details of the Chinese Room scenario are constant, at least functionally, there is scope for immense subtle variation in the different intuitive interpretations of the scenario. Thus, functionalism, at least in its stronger forms, seems to be just wrong—functional equivalence does not necessarily imply equivalence for our intuitive interpretation. There are two possibilities that follow from this: either there are special causal powers associated with different physical forms despite the apparent functional equivalence—and this seems to be Searle's view—or different physical forms affect our intuitive interpretation in different ways. I want to look at the extent to which this second explanation can account for the radical variation in different people's interpretation of the Chinese Room.

Block's 'homunculi-headed robots'

Another, related, thought experiment that highlights the same intuitive processes is that which Block uses to illustrate his argument against one kind of functionalism (Block, 1978). This is a direct ancestor of Searle's Chinese Room. Block's thought experiment is based on a robot, which is externally indistinguishable from a human body, but which internally is built of a large number of homunculi communicating with each other according to simple rules. Block's point is that, for a homunculi-headed system like this, "there is prima facie doubt whether it has any mental states at all... In Nagel's terms (Nagel, 1974), there is a prima facie doubt whether there is anything it is like to be the homunculi-headed system" (Block, 1978).

This experiment is another member of this family of thought experiments on the ascription of mentality. In many respects Block's thought experiment differs substantially from Searle's—it does, for example, offer a chance to embody cognitive theories in more concrete way than in the Chinese Room, and this gives it a flexibility which allows Block to defend a kind of functionalism. Even so, the intuitions that it invokes are similar and also belong in the realm of common-sense psychology. And like Searle's thought experiment, Block's also depends on being able to take the first person point of view when thinking about the system.

There are two relevant points that Block makes from this. First, the intuitive power of an example, like this one, involving about a billion homunculi, may be due to a "'not seeing the forest for the trees' illusion" (Block, 1980); that is, there can be a strong dichotomy between the different scales—or levels—at which we look at the system, and this dichotomy invokes the effect of the structure factor in our theory of anthropomorphism in common-sense psychology, discussed in chapter 7. This dichotomy between levels, and the intuitive effect of the structure factor of anthropomorphism on thought experiments, is an important issue for the methodology of cognitive science, and one I will return to in chapter 14. Block's second point is that "brute untutored intuition tends to balk at assigning intentionality to any physical system, including Searle's beloved brains" (Block, 1980). Here he is arguing that Searle is using his knowledge of the biological nature of brains to *prescribe* rather than ascribe intentionality to some systems, simply because they are made of the right kind of stuff.

So a pattern is beginning to emerge. It seems that all sorts of thought experiments along the lines of Searle's and Block's rest on a shared body of intuitions inherited from human common-sense psychology. This means that we can use them to study these intuitions, but it also means, as Block (1980) suggests, that we need to be rather more careful about where these intuitions are coming from and what they are doing to our thought experiments. This is an important point which I'll come back to in my analysis of the methodological implications of common-sense psychology in chapter 14.

Common-sense intuition in the Chinese Room argument

There are several hints that show Searle's thought experiment has its roots in intuition. The first of these, as I've already mentioned, is that Searle insists on our taking the first person stance to the Chinese Room, or, rather, to Searle-in-the-room ("remember, in these discussions, always insist on the first person point of view", Searle, 1980). Asking us to identify with Searle-in-the-room, as we have seen, is actually appealing to our common-sense psychology, and to our intuitions about what can have mentality and what cannot.

The second hint is the general class of the thought experiment. It does, after all, look at the ascription of mentality—the self same property that is the realm of common-sense psychology. It is, therefore, almost inevitable that common-sense psychology will be invoked, at least in part, when looking at experiments of this general form. The same goes for other thought experiments which deal with the ascription of mentality, such as those of Dennett (1987; 1991) and others of Searle (e.g. Searle, 1992).

A third hint is the close connection between the Chinese Room and the Turing test. As I've shown in chapter 5, many aspects of the Turing test actually depend on the intuitions of our human common-sense psychology. It would be surprising, therefore, if Searle's thought experiment (which assumes that the room has passed the Turing test) didn't also depend on these intuitions to some extent.

Searle's dependence on intuition in the Chinese Room argument did not go unnoticed in the commentaries on his original article. He was separately criticised by Block (1980), Dennett (1980), and Pylyshyn (1980) for this reliance on intuition. Pylyshyn is most explicit on the matter: "we cannot take as sacred anyone's intuitions about such things as whether another creature has intentionality—especially when such intuitions rest (as Searle's do, by his own admission) on knowing what the creature is *made of*" (Pylyshyn, 1980, original emphasis). In attacking Searle for "anthropocentric chauvinism" he clearly sees the effect of human common-sense psychology on the intuitions the Chinese Room is intended to invoke.

Block also notes that before using arguments such as Searle's we should have a proper understanding of the effects of these intuitions, as without this understanding we cannot escape this implicit anthropocentricity. "An argument such as Searle's requires a careful examination of the source of the intuition that the argument depends on" (Block, 1980). And this thesis' examination of these intuitions add an implicit observer-relative nature to mentality—and therefore undermine Searle's central distinction between original and derived intentionality.

Searle countered the criticisms of unbridled intuition by arguing that many of the apparently intuitive aspects of the thought experiment are, in practice, just plain facts—"it is a plain fact about me that I do not speak Chinese" (Searle, 1980). So while he accepts that intuition does, in part, figure in any such argument, he dismisses any role for differences between people's intuitions in this particular thought experiment. This is not convincing—the length of the subsequent debate is a testament to that. He lets the cat partly out of the bag when answering Dennett's (1980) criticism, by suggesting that, in practice, intuitions seem to become facts when the scenario is described sufficiently explicitly. And while the intuitions of common-sense psychology can be very strong, the model I have proposed would not suggest that they could ever become facts—it would only imply that individual differences between intuitions may tend to vanish when the scenario is described sufficiently explicitly. This is, of course, not the same thing at all; and a large part of Searle's argument rests on the apparent open-endedness of his scenario—a point that was certainly not lost on Dennett (1980).

So Searle's thought experiment relies fundamentally on intuitions—intuitions which can themselves be studied as one manifestation of common-sense psychology. After all, these intuitions depend on our knowledge of what something is made of and of what it looks like—the similarity and structure factors in the theory of anthropomorphism I developed in the second part of this thesis. Similarly, many replies to Searle's argument have also depended on inverting these intuitions by modifying the thought experiment, so they also appeal to intuitions, just different ones. To investigate the thought experiment more deeply, then, I'll look at the role of intuition in these replies.

Common-sense intuition in replies to the Chinese Room argument

Now we can return to the main replies to Searle from the perspective of intuitive appeal, and look at the effects of different physical, structural, and functional forms on the intuitions that people form. One of the most important points to note here is that the intuitive appeal of the different replies is not essentially 'rational' or 'cognitive'. This is not to claim that cognitive processes are not involved, or that cognitive psychology cannot be brought to bear on the study of these proc-

esses, but it is a claim that the ascription of mental states is something that is so little understood as yet that labelling it in this way would be putting the cart before the horse. Going back to the Turing test for a moment, it is rather interesting that Penrose (1989) sees the test as a test for consciousness rather than as a test for intelligence. Searle has also made a similar interpretation, both of the Turing test and of the Chinese Room argument (Flanagan, 1993). The test is, over and above a test for intelligence, a test of whether 'there is anyone at home' in the system. It is, therefore, a test of the ascription of consciousness rather than a test 'for' consciousness itself. There is an appeal to the 'what is it like to be' (Nagel, 1974) interpretation of consciousness at the heart of the test, so in the Chinese Room and all its various interpretations, a lot hangs on the individual differences in people's ability to take the first person stance, to *identify* with the system—with the Chinese Room itself.

Intuition in the systems reply

Taking the systems reply first, I think that a beautiful example of the effects of physical form on the intuitive judgement is provided by Haugeland's reply to Searle (Haugeland, 1980). Here Haugeland shows there is potentially an isomorphism between what is going on in a human embodied brain and what is going on in the room. Hofstadter makes the intuition in the situation even clearer by adding in a kind of slippery slope argument (Hofstadter & Dennett, 1981). This uses a set of control knobs, which provide a space of possible isomorphic models, one of which is the original Chinese Room, and another of which is a human embodied brain. Hofstadter uses this to justify one version of the systems reply (Hofstadter & Dennett, 1981).

The intuitive appeal of this reply rests on a claimed functional identity between the room and a human embodied brain. Clearly it is easy to see an embodied human brain as something with the ability to think, consciousness, emotions, and everything else that comes with being human. On the other hand, it is hard to see a computer as having those same properties, so by making explicit the analogical similarities between the two, it becomes easier to ascribe those human qualities to the computer mind. The important point to take home from this is that even if we can draw a

formal analogical relationship between two systems, it does not mean that they will be seen as having the same properties through our intuitions. In practice, systems inherit some of the way we think about them from us as we look at them.

Unfortunately, arguing against this functional equivalence is both difficult and unpopular. It is easy to be tarred by the brush of carbonism when following this path. But aside from computational Church-Turing-style notions of equivalence, there is a lot more to functional equivalence, in that even functional equivalence is at least in part dependent upon the observer, and on the observer's perception of equivalence. If this is true, there can be no such thing as 'absolute' functionalism except in an operational sense (such as the Church-Turing one) although because we are similar as observers, there will often be something that looks very like functional equivalence. The difference is that our intuitions respond differently in different individuals.

So in the framework of the intuitive appeal of a system, the systems reply is a simply test of people's ability to identify with a room rather than a person. People who can accept the systems reply are able to see the room as capable of cognition, and possibly even as having consciousness, perhaps because they can suspend their knowledge of the physical and functional structures that Searle has added to the test.

Searle's reply to the systems reply, incidentally, is fairly strong. He argues that it is possible, at least in principle, for Searle-in-the-room to memorise the contents of the rules and papers in the room, and even then he can simulate intelligence without really being intelligent. In this reply, there is nothing else there in the system, and moving to the scale of the room doesn't help when we want to ascribe intelligence. But the system doesn't just include what's going on inside the room (unless we accept that representations are *only* in the head, and we don't) it also includes the Chinese speakers outside the room, and the Chinese language and society around it. These cannot be memorised by Searle-in-the-room without his becoming part of that society, and therefore coming to understand Chinese (Wittgenstein, 1953). But since Searle's memorised version of the intuitive argument doesn't affect the philosophical argument fundamentally, and yet simplifies the example, I'll use this modified version of the thought experiment for the model in the next chapter.

Searle's reply, by replacing the entire contents of the room with Searle-in-the-room, seems to amplify the intuitive diversion implicit in the reply. Since the whole contents of the room are represented by a person, it is almost impossible to resist identifying with the person rather than with the room as a whole. This shift in intuition is perhaps most clearly shown by Hofstadter's comment "which level does Searle wish us to identify with" (Hofstadter & Dennett, 1981). As I suggested in chapter 9, there is a close affinity between the effect of the structure factor in anthropomorphism and the use of levels in psychological models—an effect I'll discuss more fully in chapter 14.

Intuition in the robot reply

The second reply I'm going to look at is the robot reply, which is perhaps best argued by Boden (1988) and Harnad (1991). Again, it is important, at least for my purposes, to separate the intuitive appeal of the reply from the philosophical argument—which in Harnad's case is about grounding symbols and how direct sensory connections can provide 'real' intentionality.

In the robot reply, the context changes from understanding language to seeing. This seems to increase the ability to identify with the system, in that it seems easier to identify with something seeing than with something understanding. There are other differences, too; the description Searle uses creates a far stronger connection with the human form in general: "it would have arms and legs that enabled it to 'act', and all of this would be controlled by its computer 'brain'" (Searle, 1980). Searle creates an image of something far more human-like, both physically and in terms of its mental competences, than he did with his original room scenario.

The main intuitive point following from the robot reply is that it shows how obviously functional equivalence depends on the way the system is seen. It is brilliantly anthropomorphic! This is not meant as a criticism; it is actually a very sharp way of inverting Searle's use of a 'room' as something that is difficult to identify with, and to turn it into something that is relatively easy to identify with.[1]

[1] Again, remember that this intuitive argument parallels the philosophical one. I am not claiming that the philosophical argument rests on something as weak as anthropomorphism. Harnad's reply is philosophically relatively strong, although it too has its problems with intentionality (Hauser, 1993).

Intuition in the other minds reply

The third reply I'll look at is the other minds reply, which is less frequently adopted, although it has been pretty convincingly argued by Hauser (1993). Searle's reply to this is perhaps his weakest—he says that to make the other minds reply is simply to ignore some of the most important tools (all our intuitions, in fact) that we have for making judgements about other people's mental states, that it is, in effect, to "feign anesthesia" (Searle, 1980). Searle's argument depends for its weight on a categorical distinction between people's minds and possible computer minds, between real intentionality and 'as if' intentionality. While these classes are often so far apart that they do appear to be categorical distinctions, as I suggested in chapter 2, there is some reasonable doubt that this is the case. And as Hauser argues, these issues force Searle into the problems with other minds that lead to his "cavalier dismissal" (Hauser, 1993) of the other minds reply.

Curiously, this stance has often been better argued from a sociological position than a psychological one (e.g. Collins, 1990; Woolgar, 1985). From a sociological point of view, there are no obviously good reasons why, in principle at least, a computer mind cannot be indistinguishably similar to a person's mind. There are a various slippery slope arguments for this; for example, there are no obvious categorical distinctions (not even biological ones) which prevent machines from becoming arbitrarily close to human psychology. The importance of the other minds reply to the sociological stance perhaps shouldn't be surprising; for example, Collins (1990) "Ultimate Turing Test" can be thought of as extending the protocol for the Turing test so that it really is more like everyday social interaction. Inevitably, then, it will begin to level out the differences between human and computer minds.

Consciousness and the ascription of mentality

In the introduction to this thesis, I claimed that I did not intend this thesis to be 'about consciousness'. I stand by this; I do not particularly want to get involved in any philosophical or psychological discussions on what actually constitutes consciousness. Unfortunately, as Searle's thought experiment shows, it is not possible to avoid the problem completely. This isn't a problem that is

specific to Searle's thought experiment; at its heart it goes back to the Turing test, which is increasingly being seen not only as a test for intelligence, but as a test for consciousness (e.g. Flanagan, 1993; Michie, 1993; Penrose, 1989).

But as the discussion in chapter 5 made clear, the Turing test has its own traps, and once again, we must be sure in our interpretation of the test before we can call it a test 'for' anything. Michie (1993), for instance, takes an operational interpretation of the Turing test and yet also takes it as a test for consciousness, and is therefore pushed into a definition of "operational awareness". I have already resisted this operational interpretation in favour of Moor's (1976) inductive interpretation, where we take the Turing test as a format for gathering evidence, in this case about the consciousness of the system we are testing.

Before following this line too far, though, I want to put this through an inversion. In practice, I am using the Turing test as a format for gathering evidence not about consciousness, but about the ascription of consciousness. Instead of looking at a property of the system, I am looking at a property of the observer. I will follow this pattern of inversion more fully, and show how it touches on the fundamental nature of the Turing test, and on the general nature of intelligence, in chapter 13.

There is an important issue here, though; one that deserves fuller recognition before continuing with the main line of the thesis. It is the strikingly close connection between the ascription of intelligence, the ascription of consciousness, and common-sense psychology. If this connection is as strong as it seems, no true understanding of consciousness will ever be possible without a fuller understanding of the processes and regularities of common-sense psychology. Consciousness, and likewise intelligence, need not necessarily be objective phenomena; they might be deeply tangled with the observer's—and the scientist's—common-sense psychology. This is a similar conclusion to French's (1990) argument that the Turing test is fundamentally centred on human intelligence—but from a very different angle. Instead of suggesting that intelligence is tangled with the human biology of the person who exhibits it, this point suggests that intelligence is tangled with the human psychology—the human common-sense psychology—of the person who ascribes it.

Before moving on, I'll put one additional gloss on this. Returning to Eddy *et al.*'s results, presented in figures 7.1 and 7.2, there is a consistent discrepancy between the levels of ascribed cognitive states and ascribed subjective states. Both show a strong correlation with the similarity factor, but for a given similarity, ascribed cognitive capacities are consistently rated more highly than ascribed similarity, while ascribed subjective capacities are consistently rated lower than ascribed similarity. I don't want to draw too many conclusions from this rather tentative result, but this is a trend which does deserve further investigation. For example, it seems to show that it is harder to ascribe consciousness to something than it is to ascribe cognition to it. So while "many contemporary thinkers are willing to attribute thought to animals, but not consciousness" (Bechtel, 1992); even if they do manage to pass the Turing test eventually, the same may well be true of machines.

Conclusions

In this chapter I've discussed Searle's Chinese Room thought experiment in a lot more detail, and showed how both it and many of the replies to it depend on intuitions that properly belong in the realm of common-sense psychology. This has a number of implications.

First and foremost, and the point that I'll tackle fully in the next chapter, we can use this format inductively, in the same way that we interpreted the Turing test in chapter 5, as a way of studying these common-sense intuitions. In the next chapter, I'll show how we can use the modelling environment we developed in chapter 9 to look at these intuitions and their effects on Searle's thought experiment. That is, by modelling different intuitions in the Chinese Room we can gain insight into common-sense psychology and its role in the ascription of mentality, rather than looking into mentality itself as Searle intended.

The second point is that thought experiments can be dangerous, because their appeal to intuition means they can undercut philosophical arguments. In effect, they tell us more about our ability to ascribe intelligence under particular circumstances than they do about intelligence in any absolute sense. I'll go into this point a lot more fully in chapter 14; for now I will just say that this has significant methodological implications for cognitive science.

A third point is that Searle's thought experiment, like Turing's, is not necessarily one that should be abandoned. For example, Hayes and Ford have said "we must explicitly reject the Turing test in order to find a more mature description of our goals" (Hayes & Ford, 1995). I fundamentally disagree. It is only by taking tests like the Turing test inductively that we can find out what intelligence is, and, therefore, find out what our goals actually are. Searle's and Turing's thought experiments offer one way of studying the processes involved in the ascription of mentality. And if, as seems likely, intelligence is partly observer-relative, a deeper understanding of these processes could offer new insight into the foundations of cognitive science.

On this note, I'll turn to the next chapter, and build a model of Searle's Chinese Room, and of a kind of Turing test too, to show how sensitive this thought experiment is to common-sense psychological intuitions, and to demonstrate how these intuitions can be studied indirectly in this way.

Chapter 12
Modelling the Chinese Room

Introduction

I have already discussed the theoretical principles behind Searle's Chinese Room thought experiment in a bit of detail in chapters 5 and 11. Searle's original article introduced both a logical argument (which was, in effect, that semantics are not intrinsic to syntax) and an intuitive argument which illustrated that logical argument. As I argued in those chapters, these intuitive aspects are important because the intuitions that are generated by thought experiments like Searle's throw a new light on common-sense psychology. In this chapter, I'm going to use the modelling environment developed in chapter 9 to look at these intuitive aspects of Searle's thought experiment in much more detail.

There are several reasons why I'm using a model of the Chinese Room to illustrate this side of common-sense psychology. First, and perhaps foremost, some of the factors of anthropomorphism are not used in the false belief test, so for a more complete account of anthropomorphism a second model is required to fill in the gaps—preferably one which again doesn't require too much physical reasoning. The Chinese Room fits the bill. A second reason is that the whole approach of using "intuition pumps" (Dennett, 1980) or thought experiments can be open to question, if they are reasoned about in a way that is biased by our common-sense perceptions. (I will argue this out more fully in chapter 14.) The model in this chapter will show that it possible to predict the processes of ascription that are involved in the Chinese Room in a way that is relatively systematic, and which leads to a better understanding of the kinds of bias that might affect thought experiments. The third reason for the selection of the Chinese Room is to show that the princi-

ples which underpin the Turing test are intimately connected to common-sense psychology, and that models of common-sense psychology need to be better understood before the Turing test itself can be used with safety as a way of looking at intelligence or mentality.

This model, then, apart from its psychological content, serves as a lead into a general argument on the ascriptive nature of intelligence, which I'll put forward fully in the next chapter. For now, though, the important point which carries through into this argument is the sensitivity of intuitive thought experiments to elements normally outside the accompanying philosophical argument, such as the physical form and structure of a system. This is especially clear in the case of Searle's thought experiment.

Modelling Searle's Chinese Room

Before I move on to describe the model of intuitions in the Chinese Room in detail, I'll begin by sketching out the important parts of the model and how they represent the general format of Searle's thought experiment. Just as with the model for the false belief test, the first part of the model script describes the form and other properties of all the objects in the scenario. This time, though, the principal objects involved are Searle himself (who will play the judge in the Turing test), the Chinese Room, Alison (who this time will play the role of the control subject for the Turing test), and Searle-in-the-room, who begins the scenario hidden inside the Chinese Room.

Once again, the most complex part of the model is the representation of the subject, in this case the judge, Searle. Searle's intentional and physical stances are, in fact, exactly the same as those of the subject, Alison, in the model of the false belief test in chapters 9 and 10; therefore Searle's ascription of mentality follows the same patterns outlined in the model of anthropomorphism and common-sense psychology described in the second part of this thesis. The only real difference between the models is that Searle has some additional, rather more complex, rules which gives him a 'mental model', an ability to run through scripts in a given domain to find the right answer to some questions. This is used so Searle can compare his own answers with those of the Chinese Room, to help him decide if the room is answering questions correctly or not, and to modify his behaviour accordingly.

The two subjects in the Turing test, the Chinese Room and Alison, also have inner models. In the case of Alison, this has the same role—to act as a mental model for reasoning about the questions which Searle puts in the Turing test. (Curiously enough, this mental model is effectively endowing Alison with the ability to simulate the scenario). In the case of the Chinese Room, though, the role of the model is very different. We claim no pretence that this model really is a mental model in the same sense as for Searle and Alison; instead, the inner model for the Chinese Room is just an implementation device that Searle-in-the-room can use to generate appropriate answers.

Searle is also equipped with the tools needed to devise and carry out a simple Turing test on the Chinese Room and Alison. This consists of rules and scripts which can be used to set up a number of scenarios; these rules and scripts are defined in the files *mcrscripts.component* and *mcrdatabases.component* in Appendix C. The scenarios are put, by Searle, to Alison and to the Chinese Room, and each is followed up by a number of questions. Depending on the accuracy of the answers to these questions, Searle gradually ascribes more intentional reasoning capacity to the Chinese Room, manifesting itself as familiarity. This might seem to be a deviation from Searle's original thought experiment, but it should be remembered that part of Searle's argument depended on the Chinese Room having passed the Turing test. Searle's point is that even if the room passed the Turing test, there would be nothing there that was real intelligence. At some level, then, it is essential that Searle is capable of ascribing mentality to the Chinese Room to the point where it could pass the Turing test. I have presumed in this model that Searle was not prepared to ascribe mentality to the room before the Turing test. I have also assumed that, for the duration of the Turing test, Searle is not aware of the existence of Searle-in-the-room; otherwise, with Searle at least, the chances are that the Chinese Room would never have passed the Turing test in the first place.

In this model, at first all that Searle is aware of in the world is the existence of Alison and the Chinese Room. At some point after the Turing test, though, Searle-in-the-room, who was hidden inside the Chinese Room all along, is revealed. This changes Searle's knowledge about the structure of the Chinese Room. It is this change in structural knowledge that apparently influences Searle's judgement, and leads to his failing to ascribe mentality to the room. This effect of the

structure factor consists of two elements; the structural knowledge itself plays one role, but also, because one of the structural elements is another person, Searle will tend to ascribe mentality to Searle-in-the-room in preference to the Chinese Room. Of course, this can only happen after Searle becomes aware of Searle-in-the-room. This transition can be found in Searle's article: "as soon as we knew that the behaviour was the result of a formal program, and that the actual causal properties of the physical substance were irrelevant we would abandon the assumption of intentionality" (Searle, 1980).

The complete model, then, exercises two factors which were not really significant in the model of the false belief test; namely, the familiarity which is acquired through carrying out extended interaction in the form of a Turing test, and the structure knowledge which is revealed when Searle finds out about Searle-in-the-room. These factors complete the model of anthropomorphism in this thesis.

The structure of the model is broadly similar to that of the false belief test. The script begins by defining the all objects, with their forms and rules; this is shown in figure 12.1. Once this is done, the objects are introduced to the model, and Searle-in-the-room is hidden inside the Chinese Room. Then we can introduce Searle to the scenario; he will remain unaware of the existence of Searle-in-the-room. At this point, we can test Searle's ability to ascribe mentality to Alison and to the Chinese Room. The next step is for Searle to carry out a Turing test, throughout which the Chinese Room will answer the questions just as well as Alison, behaving so as to validate taking the intentional stance to the room. Then, once again, we can evaluate Searle's ability to ascribe mentality to the participants. Finally, we take Searle-in-the-room out of the Chinese Room and put him back in again, all in front of Searle, who will then be aware of the existence of Searle-in-the-room[1]. This will affect Searle's ability to ascribe mentality, so we end the scenario by testing whether he can take the intentional stance to Alison and the Chinese Room one last time. The script for all these actions is shown in figure 12.2.

[1] Note that this 'trick' only affects Searle's knowledge of the structure of the Chinese Room. Revealing Searle-in-the-room in this way just tells Searle that there was something inside the room during the Turing test. Searle was previously unaware of this, because Searle-in-the-room was hidden inside the room before Searle was introduced to the scenario (see figure 12.2). The model of Haugeland's Room (the neural reply) uses the same technique to tell Searle that the room is filled with neurons.

The model Turing test is run in the domain of false belief tests, rather similar to that in chapter 10. There are a number of objects allowed, some of which, as containers (such as a box), can be used to hide others (such as a marble). There are also a number of dolls (such as Anne and Sally) which can enter or leave the room at different times. The Turing test is set as a number of randomly generated and different false belief scenarios like this, each followed up with questions, again like those in Baron-Cohen *et al.*'s false belief test in chapter 10. Each scenario is put to Alison and the Chinese Room separately, and then followed up with the same questions. The more questions that each subject gets right, the more they are ascribed competence in the intentional stance, adding to the ratings for the familiarity component. The complete trace for the basic model of the Chinese Room is shown in the file *mcr.trace* in Appendix B.

Many of the rules and scripts used by this model are only important as an implementation and do not pretend to contain anything of psychological relevance, so I won't describe them in any detail here. These include, for example, the rules which are used to generate the scripts for the Turing test used in the model, and the rules which generate the scripts for the false belief tests which make it up.

```
;;; Define alison with the form of a person, and an object database which provides her with a 'mental
;;; model' she can use to answer questions in the Turing tests run by Searle. See the file
;;; mcrdatabases.component in Appendix C for the definition of this object database.

(model-object alison :form person
  :database modelled-alison-object-database)

;;; Define the chinese-room with a different form to alison, but otherwise behaving identically. Again,
;;; see mcrdatabases.component in Appendix C for its object database.

(model-object chinese-room :form room
  :database modelled-room-object-database)

;;; Define searle-in-the-room with the form of a person, but without any object database as we don't
;;; do anything with searle-in-the-room's reasoning.

(model-object searle-in-the-room :form person)

;;; Finally, define searle himself, with the standard physical and intentional stances. The only difference
;;; from the false belief test is that searle has an object database which provides all the rules for
;;; generating and running Turing test scripts. This is defined in mcrdatabases.component and
;;; mcrscripts.component in Appendix C.

(model-object searle :form person
  :stances (basic-intentional-stance
            basic-physical-stance)
  :database searle-object-database)
```

Figure 12.1. Objects for the Chinese Room model

Finally, for pragmatic reasons, the familiarity effect in the Turing test has been dramatically speeded up, by giving the familiarity component of anthropomorphism a fairly high accumulation ramp. In practice, a test good enough to convince Searle would probably have to take quite a long time—but once again, this is an area where individual differences would be significant; Turing tests have, in the literature, been given durations anywhere between five minutes (Turing, 1950) and a lifetime (Harnad, 1992).

```
;;; To begin, set up the situation. This means putting everything except Searle in the room, and hiding
;;; Searle-in-the-room in the room. Then we can put Searle into the room.

(tell-model (put-in alison room))
(tell-model (put-in chinese-room room))

(tell-model (put-in searle-in-the-room room))
(tell-model (put-in searle-in-the-room chinese-room))

(tell-model (put-in searle room))

;;; At this point we can ask Searle whether Alison or the Chinese Room actually can believe anything.
;;; The status depends not on the returned environment, but on whether or not Searle can take the
;;; intentional stance to the object.

(ask-object-if searle (believes alison (inside ?something ?where)))
(ask-object-if searle (believes chinese-room (inside ?something ?where)))

;;; This done, we ask Searle to run a Turing test on the Chinese Room, comparing it to Alison. The
;;; Turing test script generation is shown in full in mcrdatabases.component in Appendix C.

(ask-object-to searle (do-turing-test))

;;; To find out the result of the test, once again we ask whether Searle can take the intentional stance
;;; to the Chinese Room, when compared to Alison. The Chinese Room passes the test if Searle can
;;; take the intentional stance to it.

(ask-object-if searle (believes alison (inside ?something ?where)))
(ask-object-if searle (believes chinese-room (inside ?something ?where)))

;;; Now we reveal Searle-in-the-room to Searle. Do this by taking him out of the Chinese Room and
;;; then putting him back in again. This only serves to make Searle aware of Searle-in-the-room's
;;; existence.

(tell-model (take-out searle-in-the-room chinese-room))
(tell-model (put-in searle-in-the-room chinese-room))

;;; Now Searle should ascribe mentality to Searle-in-the-room rather than to the Chinese Room. Test
;;; this by finding out, once again, whether Searle can take the right stance to the Chinese Room.

(ask-object-if searle (believes alison (inside ?something ?where)))
(ask-object-if searle (believes chinese-room (inside ?something ?where)))
(ask-object-if searle (believes searle-in-the-room (inside ?something ?where)))
```

Figure 12.2. The Chinese Room scenario script

Comparing models: Hofstadter's Room (the systems reply)

Once again, I'll make three comparisons against the basic model of Searle's intuitions regarding the Chinese Room. The first of these is a model of Hofstadter, whose intuitions about the thought experiment were diametrically opposed to those of Searle. Hofstadter's reply to Searle was in the systems reply class; that is, Hofstadter could still ascribe mentality to the room, even when he found out that inside there was only Searle-in-the-room.

In Hofstadter's Room, there are only two slight changes to the model, shown in the changed dispositional factors in figure 12.3. These affect the weights attached to the similarity and the structure factors. Hofstadter, being a functionalist of sorts, is far less inclined to be dissuaded from taking the intentional stance by the mere physical appearance or functional structure of the system. Reducing the effect of the structure factor means that Hofstadter is still able to ascribe mentality to the room even after Searle-in-the-room has been revealed.

```
;;; The fundamental differences for this model are reduced effects for the similarity and structure
;;; factors, corresponding to Hofstadter's reduced prejudice from appearance and functional structure.
;;; Compare these values to those in figure 9.6, and derived from table 9.2.

(model-stance systems-intentional-stance
  :class 'intentional
  :built-on 'basic-intentional-stance
  :disposition-energy
    (generic-function (disposition &rest keys)
      (:method ((disposition (eql 'similarity)) &key from to &allow-other-keys)
        (transform (get-similarity (object-form from) (object-form to))
          :scale -2.0 :translation 0.0))
      (:method ((disposition (eql 'structure)) &key from to &allow-other-keys)
        (transform (get-structure (object-name from) (object-name to))
          :scale 0.5 :translation 0.0))))

;;; A new definition of searle, replacing that in figure 12.1, using the new disposition weightings
;;; defined above.

(model-object searle :form person
  :stances (systems-intentional-stance
            basic-physical-stance)
  :database searle-object-database)
```

Figure 12.3. Dispositional differences for the systems reply

Comparing models: Harnad's Room (the robot reply)

The second comparison I'll make against Searle's intuitions in the Chinese Room is the robot reply, originally raised by Searle (1980) but subsequently clarified and argued by Boden (1988) and Harnad (1991), among others. Harnad's argument is considerably more subtle than the version in this model, but, just as with Searle's original thought experiment, I am focusing on the intuitive appeal of the reply rather than its philosophical structure.

The robot reply assumes a functionally isomorphic model to Searle's Chinese Room, but places Searle-in-the-room in a slightly different context—in the head of a robot. By this, the physical form of the room effectively changes to be that of a robot. This creates an additional anthropomorphic effect which goes some way to counterbalancing the effect of the structure factor caused by knowing about Searle-in-the-room. This time, the dispositional weight for the similarity effect doesn't need to be touched, simply changing the physical form of the room will increase the tendency to take the intentional stance; this is shown in figure 12.4. The result is that, once again, it is possible to ascribe mentality to the room after Searle-in-the-room has been revealed, although this time for a very different reason.

Comparing models: Haugeland's Room (the neural reply)

The last reply I'll look at is Haugeland's (1980). Haugeland's reply to the Chinese Room is precisely the kind of intuitive inversion recommended by Dennett (1980). Haugeland changes the structure of the room so that, instead of there being slips of paper inside, there are lots of neurons all connected together—in fact, a real brain inside a real head, except that all the connections

```
;;; The fundamental difference for this model is that the physical form of the Chinese Room has
;;; changed.  It is no longer a room, but should now have a form resembling that of a human.

;;; Define the chinese-room with a different form to before, as a robot; compare with figure 12.1.
;;; Only the appearance of the chinese-room is changed, its behaviour stays the same as before.

(model-object chinese-room :form robot
  :database modelled-room-object-database)
```

Figure 12.4. Object description for the robot reply

between the neurons are blocked. Searle-in-the-room is replaced by a super-fast demon, who lives inside the head and 'tickles' the neurons, using an instruction book to anticipate all the connections and intervening to overcome the blockages between them. In this way Haugeland manages to construct an alternative model of the Chinese Room which is functionally isomorphic to the original, but which once again exploits the effect of the structure factor. This time, the structure is as close to that of a real biological brain as it is possible to achieve within the framework of the original thought experiment. With regards to the intuition, the demon is, of course, not quite the same as a human being very small and very fast, but this time there is no way for Searle to reply as he does to the systems reply—there is no way that the demon can 'internalise' the neurons and their connections.

The model for Haugeland's reply is rather too complicated to present here in full detail, but it is shown complete, with a corresponding trace of the model, in the files *mcrneural.model* and *mcrneural.trace* in Appendix B. This complexity is mainly due to having to fill up the room with a significant number of neurons. However, we can model the structure scale effect without having to introduce billions of neurons—the main part of the structure scale effect is simply to inhibit anthropomorphism for structural or functional models which are too complex to have their behaviour predicted by us humans. Even relatively small networks can be complex enough for that.

Back to the thought experiment

Equipped with these results, we can now go back to Searle's article and look at the effect that the intuitions raised by the thought experiment have on the original argument. Perhaps the first and most important point is that the intuitions raised by the thought experiment have little or nothing to do with language understanding in practice, which was the original target of Searle's thought experiment. In practice, the intuitions raised by the thought experiment are those of common-sense psychology, and deal with the general ascription of mentality to the room or to Searle-in-the-room. This ascription is not specific to the coinage of understanding or intentionality as Searle suggests, and this is why the same thought experiment extends so easily to consciousness.

A second point, however, seems to back up Searle's intuitions. There really does seem to be some difference between different physical forms, in that we do find it easier to ascribe mentality to some forms than to others. This is not, as Searle suggests, because of any specific para-functional property of the material that goes to make up these forms—it is because we, as human ascribers, simply find it easier to see some forms as having mental states than others. Functionalism, as a philosophical position, may not be false, but people are not functionalists as far as common-sense psychology is concerned. This should make us cautious about building large theories on functionalism.

There are other ways that the intuitions underpinning Searle's argument can be somewhat defused through the model of common-sense psychology presented here. The replies I have described only depend on the factors of similarity and structure. The other factors can also be used. Familiarity can be used by running, in effect, a much longer Turing test; this, too, might have the effect of still allowing Searle to ascribe mentality after Searle-in-the-room has been revealed. Animation and context cannot so easily be used in the classic Chinese Room framework, without more substantial changes to the thought experiment; these changes are, of course, possible, without having any fundamental effect on the functional roles in the scenario.

But perhaps the most interesting alternative is to explore the effect of the interaction medium. If, for example, the Turing test was carried out not using a teletype, but, say, a telephone—a modality which we conventionally associate far more with people—the result, once again, may be different. Moor has proposed a 'virtual reality' version of the Turing test, and this too could be an interesting development for research on common-sense psychology in the test, because it would seem to allow some of the other factors involved in the ascription of mentality, similarity for instance, to have fuller rein.

So, although Searle's attack on strong artificial intelligence is not watertight, it does make some important points—points that perhaps Searle did not foresee—that do attack some of the central tenets of strong artificial intelligence. These points lie in the differences between people's interpretations and intuitions rather than in the philosophical argument itself. Even finding that different people have radically different intuitions when faced with this same thought experiment shows

how much the ascription of intelligence depends on the observer rather than on absolute properties of the system. And the characteristics of the observer that influence the ascription of intelligence include the factors of anthropomorphism discussed in chapter 7; similarity, familiarity, animation, structure, and so on.

I do not want to suggest, as does Hayes, that cognitive science "consists of a careful and detailed explanation of what's really silly about Searle's Chinese Room argument" (Lucas & Hayes, 1982). This is far too harsh a criticism; instead, I want to suggest that a careful and detailed examination of the role of intuitions in thought experiments like this can help to study the behaviour of those intuitions in their own right.

Discussion

The very idea that the Searle's intuitions can be modelled computationally might itself seem faintly ludicrous, but I do believe that this model of the Chinese Room does throw up some significant points. These points are important not only to this particular thought experiment, but relate more deeply to the Turing test and to the nature of intelligence in general. I will return to some of the deeper implications of the model, and of the effects of common-sense psychology on the ascription of intelligence, in the next chapter, but first, I'll review the specific points raised by this model.

First, it really does seem that Searle's intuitions can, admittedly in only a very shallow way, be modelled using this environment. The model clearly shows the ascription of intelligence being withdrawn as soon as Searle becomes aware of Searle-in-the-room, and this does seem to be precisely the effect this knowledge has on Searle's intuitions. It is important to remember when considering this, though, that the factors in the model of anthropomorphism are still only represented in a coarse and approximate manner.

And second, Searle's intuitions in the Chinese Room thought experiment do match the predicted regularities in our model of anthropomorphism—this does seem to support the use of this model in areas where anthropomorphism combines with common-sense psychology. The model does seem to show that the processes which underpin anthropomorphism may be the same processes that underpin intuitions in thought experiments like Searle's.

Part Four

Implications of common-sense psychology

BLANK IN ORIGINAL

Chapter 13

Common-sense psychology and the inverted Turing test

Introduction

Already, in chapter 5, I have discussed the Turing test in some detail and shown that while it doesn't really make sense as a scientific test for intelligence, in many ways it does offer a common-sense psychological test for intelligence. Perhaps this should not be surprising; as Dennett says, "Turing's test is not just effective, it is entirely natural—this is, after all, the way we assay the intelligence of each other every day" (Dennett, 1985). Harnad also comments that it is "no less (nor more) exacting a test of having a mind than the means we already use with one another in our everyday practical solutions to the 'other minds' problem" (Harnad, 1991).

In this chapter I'm going to look at the role of common-sense psychology in the Turing test in a bit more detail. In particular, I'm going to argue that if there is a real natural faculty involved in understanding other minds, this faculty is involved on a routine basis in dealing with the psychological 'other minds' problem and in the Turing test. Certainly, more recent interpretations of the Turing test have stressed that the observer is an active participant in the test, and that intrinsic characteristics of the system may matter less (e.g. Caporael, 1986; Collins, 1990; Narayanan, 1996; Whitby, 1996). And as I suggested in chapter 5, for the most part, it seems to be the deep connection between the observer's common-sense psychology and the Turing test that makes the outcome of the test less dependent on the system's actual behaviour.

A revised version of this chapter has been published as "Naive Psychology and the Inverse Turing Test", *PSYCOLOQUY*, *7*(14), 1997.

Accepting that the Turing test is based on common-sense psychology has a number of implications. First, it suggests that the Turing test is open to a number of systematic biases, both to false positives and false negatives, to the extent that the observer's common-sense psychology influences judgement in the test. I have already argued for this in a little detail in chapter 5, but the main case for it was developed in the discussion and model of Searle's Chinese Room thought experiment in chapters 11 and 12. In this chapter I will focus on a second, no less important implication. I will argue that the extent to which the Turing test depends on common-sense psychology implies that we can use this format, inductively, as a tool to study common-sense psychology in its own right.

This has one very important consequence. If the Turing test really is like the way we ascribe minds to each other every day, the connection runs two ways. It may be that the property we call 'intelligence'—and which for a long time was assumed to be an intrinsic characteristic of a system (e.g. McCarthy & Hayes, 1969; Newell & Simon, 1976)—was actually observer-relative all along. But before I get into that particular controversy, I'll start closer to home and look first at the role of common-sense psychology in the Turing test, both in the system and in the observer.

Does a system need common-sense psychology to pass the Turing test?

One problem with using the Turing test to assess common-sense psychology is that it is conceivably possible for a system to pass the test without having any common-sense psychology. This is a variation on the standard objection to the Turing test: that it is possible for a mindless machine to pass it simply by generating the appearance of intelligence sufficiently well. Block's (1981) argument, for example, would deny the necessity of common-sense psychology just as it would deny the necessity of anything else that might be classed as part of a real intelligence.

Block's argument is based on an imagined system implemented as a finite but very large table (estimates range from 10^{50} to 10^{120} elements, where there have only been about 10^{18} seconds since the beginning of the universe). This table contains a set of appropriate next sentences, for all possible conversations up to that point. This implementation, even for short Turing tests, shows a combinatorial explosion that makes this kind of system completely impractical, but his argu-

ment remains logically correct. However, when the number of sentences is so vastly bigger than the time it took the human species to evolve, we can take it that there must be better ways to pass the test in practice.

Deciding *a priori* what is necessary and sufficient for a system to pass the Turing test is a route towards operationalism, towards defining intelligence in those terms. If we are to avoid this, we can't state unequivocally that common-sense psychology is necessary for passing the test. We can, however, argue that at least the appearance of it is necessary, for without it a system would show an inability to perceive, recognise, and respond to human mental states to an extent that would make it trivially distinguishable from any real human. On that basis, we can continue to look at the role of common-sense psychology in the Turing test.

Turning to more empirical arguments, even understanding a sonnet like "Shall I compare thee to a summer's day?" calls for common-sense psychology, let alone writing one (Turing, 1950). As Haugeland (1985) comments on Turing's imaginary dialogue: "the student has displayed not only competence with the English language, but also a passable understanding of poetry, the seasons, people's feelings, and so on". This understanding of feelings shows that the Turing test, even in its original form, does at least touch on human common-sense psychology. So even though it may be possible to pass the test without common-sense psychology in principle, it is highly improbable that the test can be passed without it in practice.

French (1990) puts forward a convincing argument that the Turing test itself is a test "not of intelligence, but of culturally-oriented human intelligence". His argument is based on the "essential inseparability of the subcognitive and cognitive levels". In effect, there isn't a clear frame around the competences that are evaluated by the test, so in principle, any aspect of human society, psychology, behaviour, or even biology may be implicitly touched on by the test. I have already discussed the problems with the 'alien intelligence hypothesis'—the idea that there is such a thing as "intelligence in general" (French, 1990) which is independent of our human point of view—in chapter 6. Here I will just comment that French's argument seems sound, and, therefore, to pass the Turing test in practice the appearance of specifically human common-sense psy-

chology will be required. If this is true, we need to study the extent to which it is possible to identify common-sense psychology in its own right, and the extent to which the Turing test can do this.

Perhaps the simplest tests for common-sense psychology are the false belief tests of Wimmer and Perner (1983) and Baron-Cohen *et al.* (1985) reviewed in chapter 5. Is it possible to test for false beliefs like this in the Turing test? Neither Wimmer and Perner's nor Baron-Cohen *et al.*'s test is principally linguistic, so at the least a certain amount of translation would be needed to set this into a Turing test. In both cases, though, there is a potential problem. The puppets in these tests can only be replaced with linguistic references to invisible others in a teletype test like Turing's, and it is an open question whether this affects the results of the test. In the visual scenario of these tests, puppets can play the roles of third parties. In a purely linguistic test, on the other hand, questions relating to those third parties move subtly from being psychological to being hypothetical. And if, as seems likely, the form and appearance of these puppets affects the ascription of beliefs and other mental states, then false belief tests like these cannot be directly translated into the purely linguistic context of the Turing test.

We are left with a few other options. First, we can look for a new test which can handle the problems of linguistic reference in common-sense psychology rather better than these existing false belief tests, even if that test happens to use other means than false belief for assessing common-sense psychology. Alternatively, we could change the modality or circumstances of the test itself, so that references to third parties can be made more easily, and judgements of ascriptions of mental states to those third parties can be evaluated. But before trying to revise the Turing test fundamentally, though, some effort should be spent first deciding whether this is actually necessary.

It is clear that the linguistic content of interaction in the Turing test is sufficient to allow questions to test for common-sense psychology. But for the test to be valid, we need to be sure that even if a common-sense psychology depends on nonlinguistic interaction it can be tested for using purely linguistic interaction. Also, common-sense psychology can be nonlinguistic, in that animals with-

out language seem capable of having common-sense psychology, to some extent at least. We are left, then, with a simple question: can common-sense psychology be evaluated purely linguistically?

The answer to this is seems to be 'yes', given enough time. After all, when we read Jane Austen, we are fully aware of the mental ascriptions even though they are transmitted purely on paper. In time, we learn just how gifted Jane Austen was in her common-sense psychology. If we suspend our disbelief and give our common-sense psychology full rein, even a completely one-way communication medium like this can create marvellous perspectives onto people's understanding and misunderstanding of one another.

But we are still left with the original problem—the tendency to false negatives and false positives in the Turing test. Even if the Turing test is ready made for evaluating common-sense psychology, that same common-sense psychology—in the observer—seems to make the test just too unreliable. Before we can really understand how ascription actually works in the Turing test, we need a better understanding of the effects of common-sense psychology on the observer's ascription of mental states.

The observer's common-sense psychology in the Turing test

As I said in chapter 5, when common-sense psychology is taken into account the observer's role in a Turing test is not as fixed as it might have seemed in Turing's original article. The psychological baggage of the observer plays an important role in the Turing test, implicitly if not explicitly. It is this, in part, which means that the test cannot be taken simply as an operational or behaviourist test with all the scientific flaws that would imply.

The question is: what is it that makes us ascribe mental qualities to systems? To be sure, the actual behaviour of the system is probably the most significant factor. But our study of anthropomorphism in the second part of this thesis shows that it is certainly not the only factor. Many other factors, appearance, complexity, and so on, also play a role. Perhaps intelligence isn't truly a property of the system at all! Perhaps it is a joint property of the system and the observer and their

context and interaction, and is, in some sense, measured by the extent to which the observer's common-sense psychology is activated by the behaviour and appearance of the system through the medium of interaction. Could this really be what we mean by 'intelligence'? For example, one of the things that people do, when trying to decide whether something else has a mind or not, is to quickly probe for its similarity to themselves and to use this as a predictor. As I've discussed in chapter 7, this is the 'other minds' argument from analogy but as a natural psychological faculty, not as an argument. The logical invalidity of the argument cannot be doubted (Hauser, 1993), but I don't claim logical validity, I just claim that, psychologically, this is what people do. Of course, this isn't deliberate or conscious action on their part, it is just the natural faculty of their own common-sense psychology acting to build their intuitions.

Turing, in his original article, proposed using a teleprinter to mediate communication between the participants. He used this as "a screen that would let through only a sample of what really mattered" (Dennett, 1985). The intended effect of this was to make the actual form of the system inaccessible to the observer: "we do not wish to penalise the machine for its inability to shine in beauty competitions" (Turing, 1950). But a screen like this has its costs and needs to be justified. If the form acts as a cue to the observer—a cue whose loss has a significant effect on the efficacy of the test—then a better screen may be needed. Others have taken Turing's position: "neither the appearance of the candidate nor any facts about biology play any role in my judgement about my human pen-pal, so there is no reason the same should not be true of my [Turing test]-indistinguishable machine pen-pal" (Harnad, 1992). I think this assertion is simply false. It is likely that there are facts about biology—common-sense biology at least—which do play a role in exactly this judgement. If the pen-pal is known to be human, that does matter to the test. Using a screen such as a teletype link doesn't just mask the form, it changes the cues given to the observer, and this can fundamentally affect the patterns of interaction. A teletype connection isn't only a screen, it changes the interaction modality. It has the effect of changing the modality into one which we normally expect to use with dumb machines, so it is easy to assume that the system under test is a dumb machine.

The effects of anthropomorphism as studied by Eddy *et al.* (1993) do indicate a natural tendency to prefer ascribing mental states to some forms over others, in a way that broadly correlates with the (false) 'evolutionary ladder'. In a sense, Eddy *et al.*'s experiment was like a very short Turing test; so quick that no true interaction is possible, yet with the physical form revealed to the observer. Under these circumstances, the perceived form does have an effect. Unfortunately, no studies have yet been carried out on the extent to which these prejudices carry forward into assessments of behaviour based on longer periods of interaction. And besides, animals and machines pose different problems for this kind of anthropomorphism. All animals, including humans, bind together a particular form and behaviour, both of which are relatively constant, that is why a probe based on similarity can work for common-sense psychology. Machines introduce a massive potential for variation in both factors. Assuming that form and behaviour can be separated seems to me to need more evidence than has yet been provided. Turing's reason, the "fairly sharp line between the physical and the intellectual capacities of a man" (Turing, 1950) is appealing when applied to animals and people: but can we really apply the same rules to machines? Is a distinction like this valid at all? In nature, if not in principle, form is bound in with behaviour, and whether we like it or not, it does have an effect on our ascription of mental states.

The inverted Turing test

As I have already shown, much, if not all, of the Turing test's power rests on the observer's common-sense psychology. It seems obvious, then, that the Turing test can be used to identify the existence of common-sense psychology, but that the strength of the test is in the role of the observer (Collins, 1990), not in the role of the system under test. This leads to an inverted Turing test, in which a system replaces the human judge: it will pass the test if it is itself unable to distinguish between two humans, or between a human and a machine of the same class as itself. It should, however, be able to discriminate between a human and a machine that could be distinguished by a normal Turing test with a human observer. In general, then, the system's powers of discrimination should be equivalent to those of a human.

Just like every other variation on the Turing test, this will undoubtedly open the door to some criticisms, and there are some immediate problems that need to be attended to in this proposal. First, as with the normal Turing test, the expected behaviour could in principle be simulated without any guarantee of validity. But while simulating an inability to discriminate is trivial, simulating an ability to discriminate which is equivalent to that of a human is a far from trivial problem. In fact, being able to distinguish seems to require all the same background and common-sense knowledge, and all the same skills, that passing the normal Turing test does.

The second problem is one of identity. If the system is asked to discriminate between a human and a system that is identical to itself, it has a special direct access which flaws the test. This is not explicitly addressed even in the standard Turing test, which potentially has a similar problem of special access. If the observer has advance knowledge of either the human or the machine participant as an individual, they can use this knowledge to bypass the discrimination with questions like 'what's your birthday?'. If we assume the Turing test prohibits this—as we normally do—then we can also assume the inverted version prohibits it. Unfortunately, there is a second kind of identity to tackle. Even if the observer doesn't have advance knowledge of either system under test as individuals, it can have knowledge through identity of form. The equivalent in the standard Turing test would be to use separated identical twins as the observer and as the control; this opens the door to specific individual knowledge which can bypass the test in questions like 'what colour are your eyes?' The point is that even in the standard Turing test, we shouldn't simply be asking 'are you physically like me?', at least, not knowingly. As far as identity is concerned, the problems of the inverted Turing test are the same as those of the standard Turing test; the participants in all cases should be chosen so the test can't be biased in this way.

A third criticism of the inverted Turing test is that it is redundant because all its power of discrimination is also available in the standard Turing test. Many other stronger variations on the Turing test are also open to this criticism (Harnad, 1991; Hauser, 1993). For instance, Dennett (1985) anticipates and rejects a rough prototype version of Harnad's (1991) Total Turing test. Harnad's proposed extension was to add robotic capacities into the Turing test, claiming that for "*total* performance indistinguishability, however, one needs *total*, not partial, performance capacity"

(Harnad, 1992, original emphasis). Dennett comments: "Turing could reply, I am asserting, that this is an utterly unnecessary addition to the test" (Dennett, 1985). Although, Dennett follows Harnad's argument, he doesn't feel that the test needs to be changed, even if robotic capabilities are required of any system which is to pass it. Dennett argues that the Turing test is already strong enough to detect robotic capabilities, when it is taken in its "unadulterated" form.

So, what would Dennett, or Turing for that matter, say about this inverted version of the Turing test? The same criticism could apply: an inverted Turing test would be an unnecessary extension, so long as common-sense psychology can be tested for within the framework of the standard Turing test. This is probably true with respect to the Turing test: that is, a critically evaluated standard Turing test without a time limit should normally be sufficient to detect the presence of a common-sense psychology. (Of course, a more complete search for intelligence might still require modally or temporally extended versions of the standard Turing test, Harnad, 1991). However, given that humans are psychologically disposed to ascribe mental states (and, therefore, common-sense psychology) to almost anything, I doubt whether this kind of critical version of the Turing test is psychologically possible without some variation in the test. I believe that the inverted Turing test does have some power over the standard one, in practice if not in principle, because it allows the psychological baggage of the system to be evaluated with respect to ascription of mental states to third parties.

Another possible objection to the inverted Turing test is that it is implicitly recursive; it describes systems in terms of the ability to recognise their own behaviour. This is true: the test is indeed recursive, but it is not an infinite regress kind of recursion, but a transactional, or temporal regress, kind of recursion. Because the participants are playing a transactional game of "guessing thoughts" (Wittgenstein, 1953) there is an inevitable mutual recursion between their actions. The point to note is that exactly the same is happening both in the normal Turing test and in normal human social interaction in the real world. And given the nature of common-sense psychology, handling of beliefs about beliefs and so on, this seems inescapable.

An inverted Turing test might seem at first to be counterintuitive, but in practice it is a simple test for the ability of a system to apply common-sense psychology in the same way that a human can. It also gains an elegance by counteracting the implicit directional bias between observer and subject in the original, and making sure interaction is evaluated in both directions, rather than just the one. This removes what Collins (1990) calls "interpretative asymmetry" from the test. Collins uses this term to describe "the one-way process in which we repair the defects of machines' interactions while machines cannot repair the defects in ours" (Collins, 1990). Collins argues that this skill at repair is learned when becoming socialised (which is why artificial intelligence systems do not have it). The inverted Turing test ensures that this ability is fully exercised.

Putting the Turing test through these inversions allows the phenomenon of common-sense psychology to be seen from both sides, that of the observer as well as that of the system. Of course, the role of the observer was implicit from the beginning, but making it explicit does emphasise this important and natural aspect of human psychology. The inverted Turing test should be seen in this light; not as a serious proposal to dismantle the critical interpretations of Turing's original version, but as an intuitive approach to ensuring that essential aspects of the behaviours the standard Turing test was originally intended to assess are properly evaluated.

Artificial intelligence and common-sense psychology

In reality, a lot of work in artificial intelligence has already been carried out which in some sense approximates common-sense psychology; much of this was reviewed in chapter 4. Most of these are built on various extended logics, and provide ways of reasoning about explicit beliefs and desires as attitudinal states (e.g. Cohen & Levesque, 1987; Cohen & Levesque, 1990; Moore, 1985). They typically allow one agent to represent and reason about the beliefs of others, to the extent that beliefs about beliefs are possible to an arbitrary order. Of course, simply using a formal logic is susceptible to a whole bag of criticisms (e.g. Harnad, 1990; Searle, 1980) but as exploratory and descriptive techniques they can help study some of the general principles involved in common-sense psychology.

But does this approach help to buy us common-sense psychology? I think not. At least, the attempt to formalise naive physics (Hayes, 1985b) also fizzled out (McDermott, 1987), and the similarity between naive physics and naive psychology, and their logical approaches, seems so very close that a purely logical approach to common-sense psychology looks doomed. Not only this, it is also open to the naturalist criticism that there needn't be any resemblance between an evolved faculty and the laws which this faculty appears to follow. There is a big difference between the logical and formalised parodies of common-sense psychology which are widespread in the artificial intelligence literature on one hand, and the actual behaviour of common-sense psychology in people on the other. Putting back common-sense psychology requires an "innately specified competence" (Clark, 1987) as well as an interactional framework good enough to allow learning on this competence. A combination of an innate structure for common-sense psychology and a interactional learning system may be the right ingredients for this kind of project.

So these logical approaches seem to be no closer to true common-sense psychology than ELIZA is to a human psychotherapist. They only provide tools with which a common-sense psychology can be analysed, described, and modelled. Artificial intelligence, though, shouldn't regard the specialised problems of common-sense psychology as necessarily outside its domain; in fact, I believe that it should rise to the challenge of this most serious of projects. As Samet (1993) puts it: "for this sort of articulation we used to look for the self-discipline that would drive an old-fashioned AI simulation". While we need to take the problems of using the Turing test as a tool to assess behaviour seriously (Colby, 1981; Collins, 1990) it does seem that computational psychology may once again become useful as a tool for studying common-sense psychology. But before following this research programme, we need to look again at the usual artificial intelligence view that beliefs and desires are all there is; a common-sense psychology founded on an ontology built only on beliefs, desires, and intentions may be fundamentally flawed. Samet: "Common sense also adds the idea that every mind is a person's mind: that beliefs and desires are states of an enduring independent self that cannot be identified with the set of beliefs and desires" (Samet, 1993).

Of course, if we were to set out to build systems which could pass the inverted Turing test, we would find it very hard. The question is: would it be worthwhile? (And here I'm ducking all the awkward ethical questions.) Certainly, it seems to require at least all the capabilities of a system which could pass the standard Turing test, but it also places an additional focus on the ability of the system to discriminate and work with the mental states of others. Where this project would lead isn't clear, but it could be fun finding out.

Conclusions

The challenges and criticisms of the Turing test have several points to make. First, common-sense psychology is a deeply ingrained and very natural human faculty, and it is an important part of the actual behaviour that is evaluated in the test. Second, the active role of the observer in the Turing is only very rarely stated explicitly. By shutting out this important source of understanding, it is possible to create deep flaws in the test. Overall, though, the Turing test may well provide a useful framework for studying the behaviour of this natural faculty of common-sense psychology.

Because the observer has this natural tendency to ascribe mental states to systems, sometimes almost without reference to their actual behaviour, an inverted version of the Turing test has been proposed, in which the emphasis is no longer on the observer's ability to discriminate between different systems, but on a system's ability to ascribe mentality with discrimination in its own right. This test seems to be at least as strong as Turing's original, but it throws a different emphasis on the behaviours that the test is intended to evaluate; an emphasis that is closely directed to the system's deployment of its own common-sense psychology.

Of course, challenging the Turing test is easy, but it doesn't necessarily move us forward in the right direction. As French (1990) puts it: "what philosophers in the field of artificial intelligence need is not simply a test for intelligence but rather a theory of intelligence". Common-sense psychology, and its associated mental paraphernalia, may not be able to provide a complete theory of intelligence. But, perhaps more importantly, it might help us to develop a theory of the Turing test. This may not seem like much of goal now, but as social conventions blur the distinctions

between our treatments of other people and of machines, analyses, like this one, of the processes underlying why we see others as minds and not just as bodies may prove to be a crucial element of our understanding of intelligence.

Chapter 14
Methodological implications

Introduction

This ideas in this thesis have significant methodological implications for cognitive science in two different ways. On the one hand, there is the point that thought experiments, in particular, can be prone to certain systematic patterns of misinterpretation and misunderstanding because of our common-sense psychology. As the model of the Chinese Room showed, intuitive arguments can be biased by our common-sense psychology, and furthermore, these biases show distinct and regular patterns. There are, therefore, methodological implications for the use of intuitive arguments in general, and for using them in psychology in particular.

On the other hand, the cognitive modelling methodology that I have used to study common-sense psychology also needs to be assessed, because if this methodology is useful, it may help future research in this field. In particular, it is important to decide whether future generations of this model could lend anything substantial to psychological research. In part this depends on the approach of cognitive modelling itself, and in part on the actual model developed within the thesis.

In this chapter, then, I'll look at these methodological points in more detail. I'll begin by looking at the effects of common-sense psychology on thought experiments, and at the ideas and issues that it raises. In particular, this study shows that using a scientific metaphor to understand mental phenomena can interact with our common-sense psychology in surprising ways—ways that don't

---

A revised version of the first part of this chapter has been published as "The Lion, the Bat, and the Wardrobe: Myths and Metaphors in Cognitive Science", S. O Nuálláin, P. McKevitt, and E. Mac Aogáin (eds.), *Two Sciences of Mind: Readings in Cognitive Science and Consciousness,* Amsterdam: John Benjamins, 1997.

always act to the advantage of the metaphor that works best in science. After this, I will look at the models developed in the third part of the thesis, and look at the advantages and disadvantages of this approach as a way of studying common-sense psychology.

Common-sense psychology and thought experiments

One of the problems with studying mental phenomena is that it is virtually impossible for us to be objective, because as observers, our own mental phenomena get in the way; it can be hard to tell what is real and what is only in the mind of the beholder. Common-sense psychology makes this especially hard, because it gives us a tendency to empathise and identify with people—and things—which we think ought to have minds. Unfortunately, one of the things we tend to identify with is the behaviour of a model for a mental system (particularly and especially when it involves consciousness) and these are exactly what sciences of the mind are trying to create.

In effect, there are two kinds of connection between the mind and a model. On the one hand, there is the role of the model as a system description providing the "substantive assumptions" (Von Eckardt, 1993) that underpin that part of cognitive science. On the other, there is an intuitive psychological appeal—a reflection of how well people can see the model as something that could be psychological. These two are intertwined threads, but they are not the same, and the intuitive appeal of the second element can significantly affect the apparent quality of the system description provided by the first. Moreover, this intuitive appeal is a psychological phenomenon, a property of the human mind, so the models and metaphors that we use are subject, to an extent, to the nature of this human natural psychology.

In this chapter I will show just how pervasive having a mind can be. It colours the whole of cognitive science, acting as a continual normalising pressure on the metaphors that we use. This is a pressure that we need to recognise, and as far as we can, counterbalance. But before we can try to counterbalance anything so insidious, we need a better understanding of the kinds of pressures that are against us. That is, in part, the role of this thesis.

Myths and metaphors in cognitive science

As I discussed in chapters 11 and 12, the argument behind Searle's thought experiment has been to an extent subverted by the apparently endemic phenomena of common-sense psychology. What Searle's thought experiment, and the analysis of it in this thesis, shows most clearly, is that the metaphor counts. If you think about the Chinese Room as a room full of bits of paper, you get very different intuitions from thinking about it as a head full of neurons, even if the two are functionally isomorphic. While this may hint that thought experiments as a rule are dangerous, I want to show other, and perhaps deeper implications.

Metaphors of an altogether grander scale come into play in the study of mind. There isn't a single science of the mind, but a variety of sciences, each with its different dominant metaphors. For one, there is the information processing metaphor with its roots in computation, logic, and symbol manipulation, but there are others, such as the connectionist metaphor with its neural analogies. As metaphors, these are ways of seeing cognitive science, but they are also ways of seeing minds. In their strongest forms, these metaphors become what Turkle (1988) calls "sustaining myths" and "provide sciences of the mind with a kind of theoretical legitimation". Sustaining myths are usually metaphors which work particularly well: for example, the computer metaphor legitimated the use of words like 'memory' well enough to demolish behaviourism and successfully advance study of the mind (Turkle, 1988), but the effects of a sustaining myth are not always positive.

Take the information processing metaphor: Searle's Chinese Room thought experiment exemplifies one variation on this theme—Searle-in-the-room is plays the role of the central processing unit of a computer reading instructions and acting on them. It is this information processing metaphor that draws most from computer science: virtual machines, serial processing, and so on. The information processing metaphor has come in for a lot of criticism over the years. After all, how can it be that the mind is like a computer? Computers are made of silicon and grey plastic, and the similarity distance between a computer and me, made of gooey stuff, protein, and grey neurons, is just too large to accept.

Perhaps this is so, but if it is, it is partly our own fault. Computers are our construction, and they were constructed in our image. "Von Neumann and Goldstine were *not* inventing the computer. They already knew what a computer was, what a computer did, and how long it took a computer to do it" (Bailey, 1992, original emphasis). Our information processing metaphor is derived from this, the human computer that was the original model for the central processing unit in today's electrical ones. We constructed the computer in our image—or, rather, in the image of one particular kind of human mathematical and logical reasoning—before we started to consider our minds in its terms. The information processing architecture was designed around a human model, and one where the roles were already clearly determined. It was possible to imagine *being* the central processing unit of a computer, acting as Searle did in the room, reading the next instruction and acting on it. Unfortunately, the strength of the metaphor is such that it is *only* possible to imagine being the central processing unit. Although throughout the system there is a finely orchestrated and synchronised system of parts all collaborating; we see the system as serial, but all along the system is parallel underneath. The serial virtual machine (and virtual machines are another kind of metaphor) supervenes on a parallel virtual machine, but psychologically it is significantly more attractive. The central processing unit acts like a psychological magnet: the strength of the metaphor draws us to that one way of seeing the system at the expense of all others.

The connectionist metaphor has a similar pervasiveness, but in a radically different way. Connectionism's sustaining myth, according to Papert (1988), is "that a deeper understanding would reveal the naiveté of such everyday analogies". In particular, the theory of eliminative materialism argues that the illusory nature of common-sense psychology (and probably the rest of psychology, too) will eventually be revealed—and they will be superseded by a deeper understanding of human neurophysiology.

In both of these cases we see two different levels of description; the distinction between the external and the internal points of view. But a need for mental coherence prevents us identifying with both levels at the same time, so if we are pushed away from one we are inevitably drawn to the other. This is the principal characteristic of the structure factor of anthropomorphism, discussed in chapter 7. It is here that we see the effect of the structure factor on science. For example,

eliminative materialism subverts its logical argument with this anthropomorphic effect: we are studying at the level of neurophysiology but rejecting identification at that level, leaving our intuitions free to associate with other levels. We identify with the whole *because* we find it so hard to identify with the complexity of interaction between the parts. And like Dennett's (1971) example of the complex chess machine, we find it easy to take the intentional stance to the system as a whole.

There is a single theme running through these cases. Parallel to the logical structure of each argument, there is an implicit test of intuitive plausibility. Ashmore (1993) describes a three player scenario involving a cat, a catflap, and an owner. In the first version the owner plays the actor, modifying the behaviour of the cat, playing the behaver, with respect to the catflap as the environment stimulus. He then rotates the players through the roles of actor, behaver, and stimulus to generate different tales of varying plausibility. In one variation, for instance, the catflap modifies the owner's behaviour through the environment in the form of the cat. Can a catflap be an actor? As Ashmore says: "The only relevant question is: does the story work? Is it plausible? By which I mean how 'comfortably' does this story distribute its particular roles and statuses" (Ashmore, 1993). As readers, the tales depend on our ability to see the players in their assigned roles.

Myths and metaphors compete for mind room in a truly Darwinian fashion. Those which are relatively successful, such as the information processing model and connectionism, can survive for many years. But this success isn't purely a matter of the corresponding theory's explanatory power for science, it also depends on its intuitive plausibility. Translating metaphors can keep explanatory power unchanged but still affect intuitive plausibility. Even if we hope that our sciences of the mind use functional rather than teleological explanations (Searle, 1992), we must always remember that any dominant metaphors within these sciences—like other sciences—are being evaluated through intuition as much as through critique.

This is an important lesson, and one that is easy to forget at unguarded moments. Even psychologists are psychological! When we interpret psychological models, we are not doing so purely on 'objective' criteria, but, because they are *psychological* models, we also ask ourselves whether that is how *we* are, inside. This is a special kind of asking—the kind of asking that appeals to the intuitions

in common-sense psychology. It is also the kind of asking that appeals to the first person point of view—and therefore to the intuition that lets us ascribe consciousness to other people, or as in this case, to models.

Consciousness and metaphors

So what of consciousness? Consciousness and anthropomorphism seem to be closely connected. When we remember Searle's thought experiment, the importance of the first person perspective highlights the connection. Flanagan (1993), for example, points out that the Chinese Room can be seen as a problem of absent consciousness, as well as absent intentionality, and Searle himself insists that the thought experiment can only be properly understood from the first person point of view.

In thought experiments like Searle's, which ask us to take the first person stance, we are asked to try to identify with the system as well as taking the point of the philosophical argument, but this identification can change the argument subversively. This effect is particularly strong with models of consciousness, which frequently ask for (and even depend on) this kind of identification. I have a tendency—irrespective of any logical argument—to believe that you are conscious which gets stronger the more I perceive you as being similar to me. The same applies to models and thought experiments: the less we are able to see similarity to a model—and this depends on us as much as the model—the more it resists identification. Of course, the model which least resists identification is simply another human; all other models seem to resist identification more than this, but to differing degrees depending on our perceptions, background, and so on. When we can't identify with the behaviour of a model, though, there is a strong psychological drift towards normalisation—to restructuring or reinterpreting the model so we *can* identify with it.

Perhaps the ultimate subversive science, then, is a science of consciousness. After all, it is in a science of consciousness that our intuitions about other minds run really deep. In this section, I am going to look at three different kinds of normalisation which are commonly found in responses to models of consciousness. These three kinds, or classes, of normalisation operate in

different ways, but they all operate through the different dispositional factors for anthropomorphism described in chapters 7 and 8, and are influenced by different properties of the system or model we are looking at.

Class 1 normalisation: 'The Similarity Fallacy'

The first, and perhaps the most powerful normalisation strategy is caused by the similarity factor of anthropomorphism. It is this effect that is responsible for constructing an 'evolutionary ladder' of consciousness, with the more conscious animals being those which are phylogenetically most similar to humans, just as Eddy *et al.*'s results showed in figure 7.2. This evolutionary ladder, often in the form of a 'scale of complexity' can be seen in Elton (1993), Chalmers (1996), and Sloman (1996). Figure 14.1 shows two different models, with a somewhat blatant similarity effect added to distinguish between them, in the form of vaguely human features added to one form. In this case, it is easier to ascribe mentality to the model in 14.1b than the one in 14.1a.

In practice, one of the strongest effects on this kind of normalisation is the use of different perceptual modalities. This shows up clearly in the robot reply to Searle's Chinese Room argument, which typifies the similarity effect in normalisation—a key feature of the robot reply is the switch from 'understanding' (of language) to 'seeing' (of objects). Direct perceptual modalities seem to increase the tendency to similarity normalisation, where non-perceptual modalities tend to decrease it.

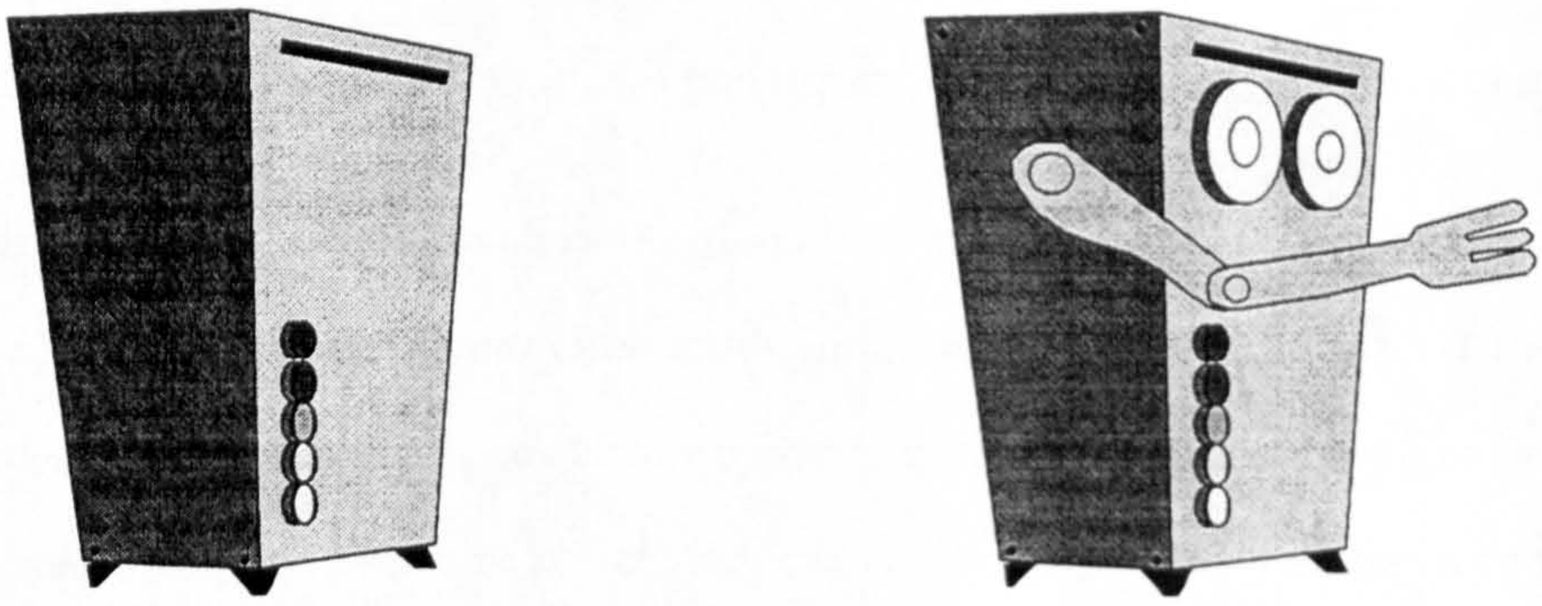

Figure 14.1. Similarity in anthropomorphism; (a) less similar, and (b), more similar

Class 2 normalisation: 'The Consciousness Box Fallacy'

In the second kind of normalisation, consciousness is taken out of the model by putting it in a separate 'consciousness box'. This box might be labelled the 'self' or 'consciousness', or it might show up in weaker forms like 'attention' or as a 'supervisory system'. In effect, this strategy makes consciousness into a homunculus, and puts everything that is generally inexplicable into this homunculus. This is the normalisation strategy that is especially common with the information processing metaphor and its use in psychology. This kind of normalisation is shown in figure 14.2, which compares external and internal views of an information processing model. In a typical information processing model, there is an active agent, the central processing unit, that reads and acts on instructions using information from mostly passive memory and perceptual systems. The central processing unit acts as the homunculus in a computer. In this kind of normalisation, it is easier to ascribe mentality to the central processing unit in figure 14.2b than to other elements in 14.2b, or to the model in 14.2a.

In this normalisation strategy, there is a distinct bias to identifying with objects with certain functional roles in the complete behaviour of a system. In the paradigm example, people find it easier to imagine themselves in the role of the central processing unit in an information processing model than they do the memory or communications modules. And even more strongly, they find it easier to imagine themselves in the role of the central processing unit than they do to see themselves as the system as a whole. This kind of normalisation can be seen very clearly in Searle's original version of the Chinese Room thought experiment, modelled in chapter 12.

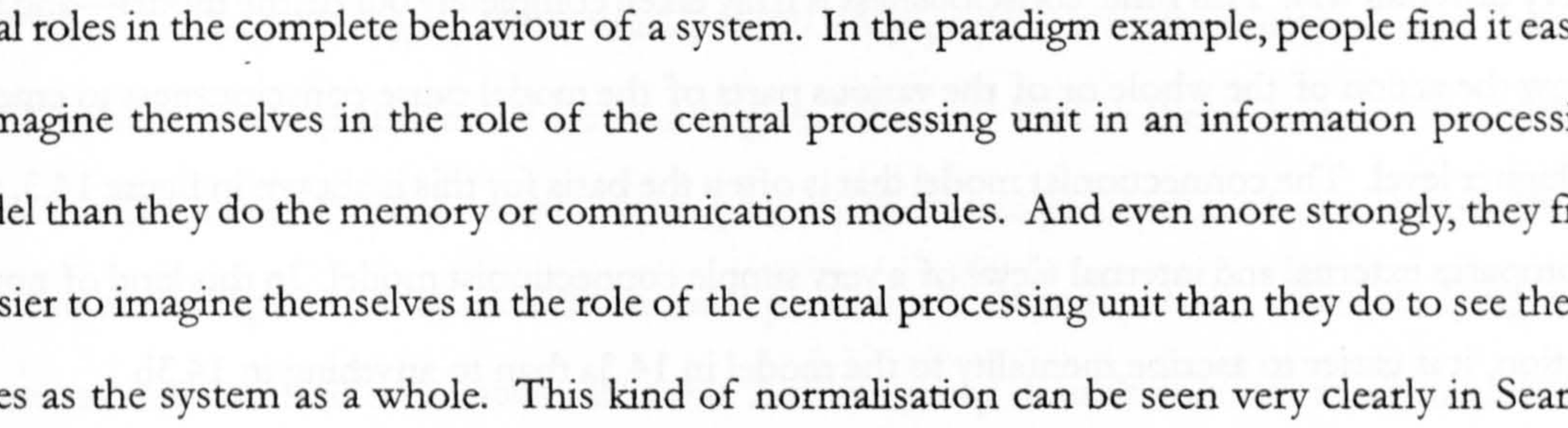

Figure 14.2. An information processing model; (a), the external view, and (b), the internal view

Note that the strength of this kind of normalisation depends substantially on what we know about the central processing unit. In effect, the more that the central processing unit approximates a human, the stronger the tendency to homuncular normalisation. And correspondingly, the less the central processing unit approximates a human, the less likely that this kind of normalisation will occur. This effect can be clearly seen in different people's thought experiments on the information processing model theme.

Examples of the consciousness box strategy can be found in Baddeley's model of working memory (as the "central executive", Baddeley & Hitch, 1974), and in Norman and Shallice's (1980) "supervisory attention system". The consciousness box strategy is also similar to the strategy that Dennett (1991) describes as the "Cartesian Theatre".

Class 3 normalisation: 'The Radical Emergence Fallacy'

In the third kind of normalising response, consciousness is again taken out of the model, but in a very different way. This time, consciousness is truly taken completely out of the model—and somehow the action of the whole or of the various parts of the model cause consciousness to emerge at a higher level. The connectionist model that is often the basis for this is shown in figure 14.3, which compares external and internal views of a very simple connectionist model. In this kind of normalisation, it is easier to ascribe mentality to the model in 14.3a than to anything in 14.3b.

This kind of radical emergence is usually coupled with a scientific strategy we can call 'level chauvinism', depicted in figure 14.4. In level chauvinism, there is a tendency to appropriate a particular phenomenon, such as consciousness, to be explained at a particular level of description. Level

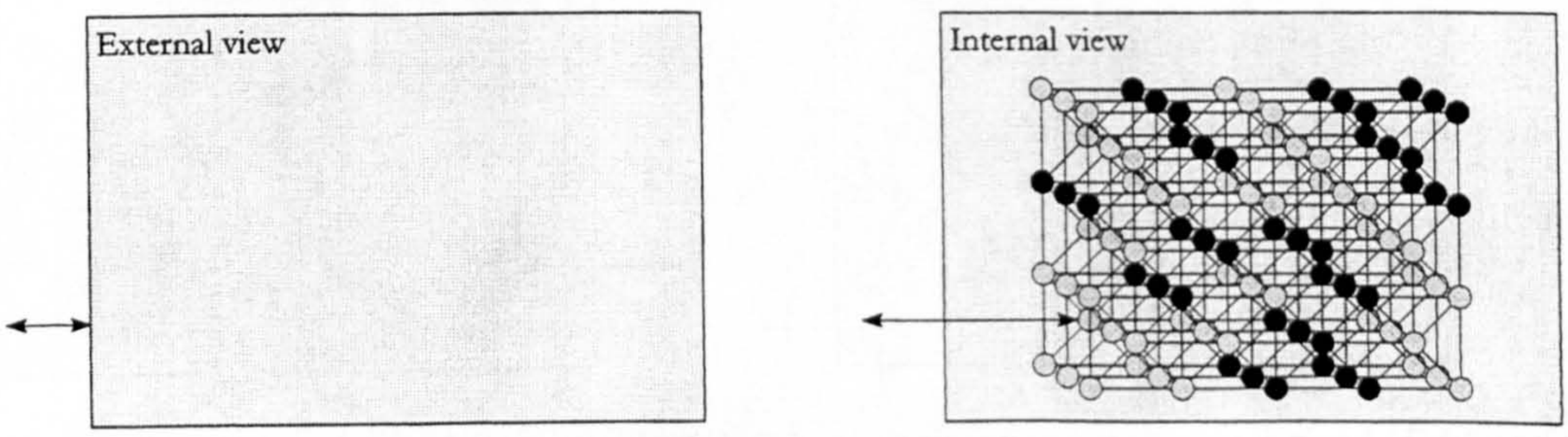

Figure 14.3. A neuroscientific model; (a), the external view, and (b), the internal view

chauvinism is dependent on the increasing specialisation of sciences into levels, described by Elias (1974). One common manifestation of level chauvinism is Churchland's (1981) 'eliminative materialism'. According to the more extreme versions of eliminative materialism, psychology is effectively redundant, because all the important parts of psychology would 'emerge' from a proper and complete neuroscience. It is one discipline saying to another: 'I can explain it, you can't'. Eliminative materialism is an example of radical emergence because it asserts that once all the small details are known, the important parts of (say) the psychology of consciousness will manifest themselves: 'as if by magic, the shopkeeper appeared'. This kind of normalisation is especially strong in connectionist approaches to artificial intelligence and quantum consciousness, as well as artificial life, and in religion. Level chauvinism, in its various forms, is an important issue for consciousness because, so far at least, nobody really has much idea about which is the right level to approach it on, or even if the whole level-based approach to science might not be inappropriate anyway (Bohm, 1980).

I am not claiming that there aren't big issues regarding the role of consciousness in a science of the mind: there are (e.g. Flanagan, 1993; Searle, 1992). My point is a methodological one: that we must be aware of the platform we are standing on when we look at these issues. Akins (1993) points out that we can "mistake our intuitive grasp of the visual perception of external events for an accurate description of internal attentional processes". It is especially when studying consciousness that we need to be aware that this intuitive grasp doesn't respond to all models and metaphors equally—and that we might be accepting or rejecting models for intuitive reasons rather than scientific ones, perhaps without even realising it.

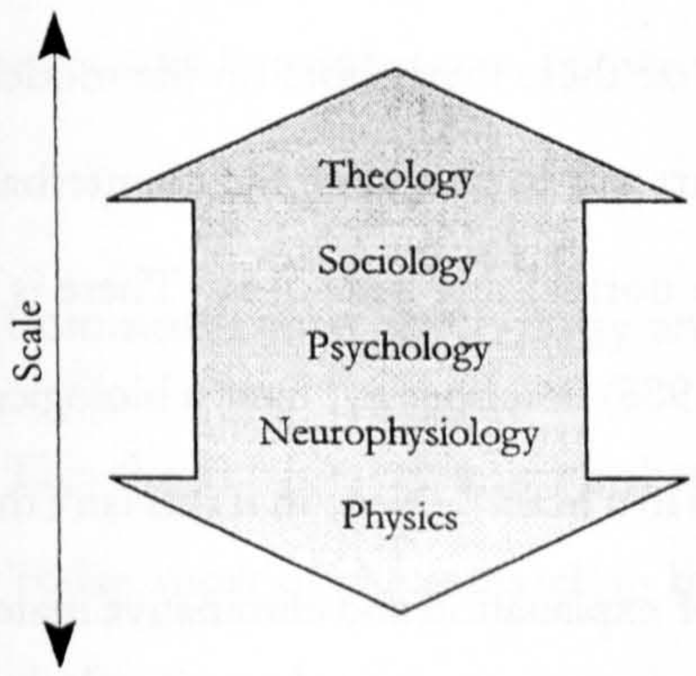

Figure 14.4. Level chauvinism

Interim summary: common-sense psychology and thought experiments

Given all these ways that our intuitions influence our understanding, it is a bit easier to see precisely why a science of consciousness is so hard to accept. Generally, these normalising strategies have led a significant number of researchers to follow different paths, either looking at completely different levels such as physics (Penrose, 1994) or neuroscience (Churchland, 1990), or giving the whole thing up as a bad job (McGinn, 1989). They all see cognitive science as incapable of solving the central problems of consciousness, although for very different reasons. Even so, it is significant that many abandon cognitive science in favour of apparently 'harder' sciences for studying consciousness. But perhaps the difference between these sciences and cognitive science is partly an illusion, a difference that has its roots in the subversive nature of cognitive science. Turkle (1988) points out that artificial intelligence (like psychoanalysis) contains a subversive component, and that "a normalising response is common to all subversive sciences of the mind". As cognitive science pushes harder at consciousness and affect, and we begin to see it becoming a little less of a mystery, there is a backlash which softens our models. The origin of the normalising response is, I believe, in the phenomenon of anthropomorphism that started us on these issues: normalising is a way of restructuring the system so that its dominant metaphor is easier for us to identify with. Introducing an executive, or attention, into a cognitive model puts the remaining mysteries into a new box, with which we can continue to identify. The normalised model merely postpones the problem. The solution to the apparent impasse is not in trying to change the science of the mind, but in understanding, recognising and resisting the normalising response.

It is extremely unlikely that we could ever form a science of the mind—or any understanding of anything for that matter—without metaphors, but we must always try to be aware of the effects of the metaphors that we use, and take into account the effects of these metaphors on the models they accompany. Even so, some metaphors do seem to go part way to providing the counterbalance which perhaps should be the correct response to these normalising pressures. There is a trend to a new group of metaphors, those which, as Turkle (1988) describes it, "have a biological aesthetic—they are the kinds of things that could be going on in a brain". Note that this isn't the same as the sustaining myth of connectionism, that of levels of explanation and eliminative materialism.

The point behind the biological metaphor is that it too appeals to our intuitions, but this time to the rather more complex intuitions that come out of common-sense biology. Unlike the rather mechanical models common in cognitive science, these go with the grain of common-sense psychology—after all, as Carey (1985) argues, common-sense biology probably emerges from common-sense psychology in the first place. The biological metaphor is not in any sense *outside* our intuitions, but it does provide another kind of hook we can hang them on. Perhaps this is why, for example, Haugeland's neural reply to Searle's Chinese Room, which constructs a neurological isomorph of Searle's scenario, manages to be so much more convincing than Searle's original. Surprisingly enough, biology matters—even if it only matters to ourselves as ascribers of mentality rather than to the system which is invoking this ascription.

So although I believe in the biological metaphor, and in evolutionary psychology, and that following Darwin, we really can learn about human mental life from animal mental life, we still need to tread with extreme caution, especially in the use of thought experiments. The new metaphor, whose slogan might be 'people are animals', has new opportunities but also new dangers. A true science of the mind emerges as the competition between these myths and metaphors drives each into refinement, and into the creation of new metaphors. I don't want argue for a single metaphor: following the Darwinian meta-metaphor, that would be a kind of scientific eugenics; I believe that the best opportunity for cognitive science lies in the natural plurality that already exists. The error is to confuse this plurality with a crisis in the discipline. The psychological and sociological pressures for normalisation will not go away, and nor will our anthropomorphic tendency to see metaphors from inappropriate points of view, but if we recognise these issues we have overcome the worst part of the problem. The role of intuitions and metaphors in cognitive science is an important one, but we must remember it isn't always under our control.

Common-sense psychology and computational modelling

This thesis has taken a somewhat unconventional route to studying common-sense psychology. Today, most of the research in this area is carried out under the banner of developmental psychology's work on theory of mind, which has undergone rapid growth since the early 1980s. But

instead of carrying out empirical research on small children, I have taken the very different strategy of computational modelling. I'll come back to the advantages and disadvantages of the modelling approach in a moment.

But this is not the first research in the field to use computational modelling. Shultz's (1991) model of the common-sense psychology of agency used the same approach. This was, perhaps, the first attempt to draw on empirical psychological research as a source for computational models of common-sense psychology, but as I have described in chapter 4, common-sense psychology also has a respectable—although completely separate—heritage in artificial intelligence. This has, however, traditionally been based on a model of common-sense psychology that was locked into a framework of explicitly represented beliefs, desires, and intentions—and even then it managed to concentrate almost entirely on beliefs. Phenomena such as anthropomorphism have been almost completely ignored. As I've mentioned, the only clear examples of computational models in this kind of field are Shultz's model already described in chapter 4, and the model presented in the second and third parts of this thesis.

Advantages of computational modelling

The general advantages of cognitive modelling have been put forward by far more experienced and skilled researchers than I, over the years (e.g. Miller, 1981; Newell, 1990; Pylyshyn, 1989). And some computational models have proved themselves to have real predictive power about human psychological behaviour. Computational modelling is useful because it complements other approaches to psychology. Most psychological research is analytic; that is, it looks at behavioural and other evidence, and uses that to make deductions about the mechanisms which might underpin the human mind. Computational modelling, on the other hand, is synthetic; it builds models of the human mind, and then looks at the discrepancies between the behaviour of the model and a real human mind, and from these discrepancies makes deductions about the mechanisms which might underpin the human mind. These methodologies are clearly complementary rather than contradictory.

One big and important distinction that should be made, if it isn't already clear, is that the methodology of computational modelling is not bound to the information processing metaphor. The most obvious counterexamples are neural networks, which are used in the same methodology, but, almost without exception, do not necessarily adopt the assumption of information processing. Although the methodology isn't necessarily committed to a single metaphor for the mind, all computational models offer one important advantage; they apparently provide unambiguous descriptions which can be compared to real human minds, by anybody in the academic community, and can be used as a kind of psychological 'gold standard' against which behaviour can be compared in some objective sense.

Computational modelling also introduces a discipline which forces clarification of existing psychological models. Often, a psychological model, when written on paper, may look complete and consistent; it is only when you try and turn this description into a working model that all sorts of gaps in the original description become apparent. This relates to a problem I mentioned earlier in this chapter, discussing the 'consciousness box fallacy' in models. According to this theory, there is something in us that pushes our models to contain black boxes we can ascribe consciousness to. Only when we come to try and build or simulate these models do these black boxes become revealed as glory holes for phenomena that don't fit the model.

Finally, computational modelling has special advantages for modelling individual differences. Simon (1975), for example, showed that the same problem solving behaviour can be generated by very different strategies in different individuals, and used the General Problem Solver to describe these differences in rules. This makes it easier to compare the different strategies, and to design experiments which can distinguish between them. The model in this thesis also shows this character; both in the false belief test and in the Chinese Room, individual differences can play in an important role. Computational modelling can be a useful tool for studying these individual differences.

Turning to the special advantages of the models developed in this thesis, this is an area which computational modelling has rather neglected. Apart from Shultz's (1988; 1991) models, there has been very little work in this field. Samet (1993) contributes a helpful critique of the methodology of research on common-sense psychology. His point is that common-sense psychology, in

itself, has never really been properly described in an unambiguous way, a way that can provide the kind of standard that is needed by an adequate theory. A theory, as we have seen, is essentially validated by its predictive utility in practice, and, Samet's criticism is that common-sense psychology is so weakly described that the theory theory—the theory of people's theory of mind—has little real predictive strength. To help provide this quality of theoretical description, Samet argues that we need "something that will work, something that will take inputs and give us reliable outputs" (Samet, 1993) and explicitly suggests that artificial intelligence could provide simulations as part of these descriptions. This is one role that I intend for the models in this thesis. Common-sense psychology is an area where there are extremely subtle distinctions, both conceptual and terminological, and accordingly, there are many public and substantial disagreements about what constitutes metarepresentation, simulation, a theory, a belief, and so on. Common-sense interpretation of these words is, in great measure, what has led to these disagreements in the first place. There is, then, a real and substantial role for a system that can form a context within which people can point to model components and use them as some common ground for these terms.

Disadvantages of computational modelling

Computational modelling, though, is not a panacea. It still needs the strand of empirical research to provide comparisons against its models; it can't simply be left to its own devices. First off, computational modelling is not something that just anybody can do. It is a skill, and for many of the more complex models, such as SOAR (Newell, 1990) and ACT* (Anderson, 1993), these skills are such that only a few people can actually use them. This isn't only because they often require advanced programming skills; as I mentioned before, a lot depends on people's ability to connect model elements to psychological qualities—this isn't always easy.

I have already alluded to the second problem in the previous section, where I suggested that computational models "*apparently* provide unambiguous descriptions"; the use of the word 'apparently' was intentional. Computational models are not really unambiguous, although they are usu-

ally a lot less ambiguous than other kinds of model. The model in this thesis fits in with this pattern, its relative unambiguity is one of its advantages, but there are areas where it remains ambiguous, such as in its connection of psychological and physical reasoning.

A third problem with computational modelling is that, almost without exception, the account of a computational model is descriptive rather than causal—it says what the behaviour is but manages to skim over how it is achieved. (Neural networks, by the way, are often equally good at skimming over this.) I have not, in this thesis, claimed that this model provides a causal account at any level, in fact, in chapter 2, I argued specifically against its offering a causal account. This is, by the way, not a problem that is unique to computational models; most psychological models that are written and communicated to other people are, in practice, descriptive rather than causal. On the other hand, until we know what common-sense psychology is, it is premature to worry about how it is implemented.

Further use of the model

The modelling framework that I have built and used in the third part of this thesis can be taken further. It can, for example, be used to explore more sophisticated experiments than the simple false belief test modelled in chapter 10. Research in the psychological theory of mind community, for example, has recently exploded into interest on a much larger scale than before. There is, therefore, a very large and rich body of empirical results beginning to accumulate in the psychological literature. Models may have an important role to play here, connecting and correlating this empirical data.

There are, however, a number of significant limitations to this model. First, the model is very shallow. The kinds of ascription involved are very limited—at present, the system only ascribes beliefs to self and others, and general mentality, no other mental states. Within this thesis, this is not a serious problem, because the main emphasis is on the distinction between agents and objects, and on the connection between intentional and physical reasoning. Neither of these different kinds of reasoning needed to be represented in any degree of richness. Further development here, then, will be in the richness of the models of intentional and physical reasoning. A second

limitation is that there are a number of sections which are psychologically implausible. These are particularly evident in some of the dispositional factors in anthropomorphism, which is, perhaps, the least well studied component of this model of common-sense psychology. More empirical results are badly needed to clarify the actual structure of anthropomorphism.

Another area where the model is implausible is that the interaction between the different dispositional factors is completely ignored—they are simply assumed not to interact. This is certainly an oversimplification. There is, for example, some interaction between the context and the structure factors. It is this interaction that leads to the 'radical emergence' strategy described earlier in this chapter—by switching between the level they focus on according to the context, people can ignore some of the effects of the structure factor.

Secondly, all the factors in the model of anthropomorphism have been oversimplified. Taking, for example, the model of similarity: for the sake of making the model tractable, I have adopted a feature-based description—a description that makes similarity metrics relatively easy to calculate with the techniques adopted from numerical taxonomy. In practice, these oversimplifications are partly down to the scale of the model, but also, and more significantly, due to the lack of knowledge in this area. The phenomenon of anthropomorphism is, in practice, so little studied that more complete models of these effects cannot yet be built.

But despite all these limitations, I believe that there are two methodological applications for this model. First, and perhaps foremost, it can, like other computational models, be gradually extended into a more complete theory of common-sense psychology. Eventually, perhaps, it will be complete enough to provide some kind of a methodological counterpoint to experimental studies in the developmental psychology community. Even in the interim, though, simply developing the model can help to find the gaps and inconsistencies in the differing theories. Indeed, developing the model in chapter 9 has already helped to do this, by showing how little was understood about the processes involved in distinguishing between things which should be ascribed mental states and things which shouldn't. And even a small model like this one can highlight—as shown in chapter 10—the subtleties of the differences between theories in some parts of common-sense psychology, such as the false belief test.

The second, less obvious but perhaps equally important, application of a completed version of this model, would be as a grounding for thought experiments such as those of Searle. I have already discussed in considerable detail how far-reaching the effects of common-sense psychology can be on thought experiments which involve the ascription of mentality. Unlike others (e.g. Lucas & Hayes, 1982) I do not want to dismiss the whole idea of thought experiments because of this intuitive contamination, putting everything into argument instead—I am not sure this is even possible. I believe that thought experiments can be a very useful tool for studying intuitions, but that we need to take time and care when reflecting on these intuitions. Models, like the one of the Chinese Room in chapter 12, can serve to provide exactly that reflection. It might be possible that in the long term a model of this kind can be used to provide some common ground for intuitions—perhaps by exploring the individual differences in people's intuitions about the same model. This common ground can then be used as a point from which to explore variations in the model and their effect on our intuitions.

So what future might this model have? Within this basic framework there are a number of extensions to the model that will be needed to build a more complete model of common-sense psychology. Perhaps the most important of these is a component which ascribes desires and intentions, and other dispositional mental states, to other agents, but there are others: a clearer connection between common-sense psychology and common-sense physics, for example, and a more sophisticated set of observation axioms, perhaps derived from Davis (1988). And while this model already incorporates many characteristics of modal logics (e.g. referential opacity)—indeed this is probably inevitable (Leslie, 1988)—future empirical study of the correspondence between these logical properties and actual common-sense psychology may also improve the scope of the model.

The idea that this model could, in the long run, become something analogous to SOAR, a kind of 'unified theory of intuition', might seem rather far-fetched, but it is not beyond the realms of possibility. Even so, models like this should not be regarded as in any sense contradictory, or alternative to, models of cognition; they can just help to make sure the ground is safe for them methodologically.

Chapter 15
Conclusions

The story so far

This thesis began as an interdisciplinary enterprise; as an attempt to bring together the different aspects of common-sense psychology in philosophy, psychology, and artificial intelligence into a more coherent body, by connecting their similarities and emphasising their differences. It turns out that once the terminological differences are dealt with, there are far more similarities than there are differences.

There are a number of common themes which show up in all disciplines. First, there is a tension between 'theory' and 'simulation' views of common-sense psychology, which manifests itself in all three disciplines under different names (in philosophy, projective as opposed to normative language; in artificial intelligence, heuristic as opposed to hypothetical reasoning). Second, there are several different theories of the actual form the constitutive elements of common-sense psychology take, for instance, the 'sentential' view shows up in Fodor's (1985) philosophy, Leslie's (1987) theory of mind, and in Cohen and Levesque's (1990) logic of knowledge and belief.

This thesis has presented a model for common-sense psychology, in two different forms. First, presented in the second part of the thesis, there is a descriptive theory of common-sense psychology, which builds on existing models of common-sense psychology, and integrates them with a new theory of anthropomorphism. Second, in the third part of the thesis, I took these ideas further and developed this theory into a computational model of anthropomorphism in common-sense psychology, which I then used to evaluate the theory and study the effects of anthropomorphism on people's everyday ascription of mental states.

In this last chapter, I will return to the theoretical background to common-sense psychology set in the review chapters, and look at the contributions made by the theory and model of anthropomorphism in common-sense psychology developed in the main body of the thesis. But because research on common-sense psychology goes on in several different disciplines, there are different perspectives on the issues. I'll return to each of these contributing disciplines in turn—philosophy, psychology, and artificial intelligence—and review the implications of the theory and the model presented in this thesis.

Implications for philosophy

Common-sense psychology is important to philosophy. The issues revolving around it include some of the deepest problems in the field—the mind-body problem, the other minds problem, the nature of intentionality, and the nature of consciousness. Nowhere in this thesis would I pretend that common-sense psychology has the complete answer to any of these deep questions, but I still believe that it does contribute to an understanding of the issues they raise. There are strong common-sense psychological themes woven not only into the practice of philosophy today, but into the philosophical problems that we have inherited from yesterday. These may become clearer, given a proper theory of common-sense psychology; on the other hand, they may not, but I suspect that at the worst, a better understanding of common-sense psychology will be needed to remove some of the mud from the waters that cover these problems. Again, common-sense psychology makes both theoretical and methodological points in philosophy.

Theoretically, most of the issues raised spring from a single central claim, that common-sense psychology is central to how people see minds, both their own and other people's. Since it is a natural evolved faculty, rather than a theory in any scientific sense, one person's understanding of other people's minds will be subject to all sorts of subtle effects—effects which are tell-tale signs of the evolutionary history behind common-sense psychology. This sets common-sense psychology as a natural faculty which cannot, therefore, be eliminated from science (Churchland, 1981), and which is richer and more complex than a representational theory of mind (Fodor, 1985).

In this thesis, I have presented a modified and extended version of Dennett's (1971) theory of intentional systems, and used this to make a number of important theoretical claims. First, Dennett's "rationality assumption" has nothing particular to do with rationality; it seems to be better described as a complex of dispositions that influence when people take the intentional stance to something. In practice, the closest approximation to the rationality assumption we have seems to be the psychological phenomenon of anthropomorphism, discussed in chapter 7. Anthropomorphism, as part of people's natural tendency to ascribe minds to other people and to things, is right at the heart of common-sense psychology—and right at the heart of this thesis.

Methodologically, common-sense psychology raises important issues about models, metaphors, and thought experiments. The theory developed in this thesis, in chapters 5, 13, and 14, and the model of Searle's Chinese Room thought experiment in chapters 11 and 12, show that, in practice, much of people's reasoning about models and metaphors is governed by the intuitions of common-sense psychology. For proper understanding of models and metaphors, then, it is important to have a good account of the biases and individual differences which can come out of human common-sense psychology. Although the theory and model in this thesis is only partially complete, it does go some way to providing this account.

Implications for psychology

In the realm of psychology, I have proposed a new theory and model of common-sense psychology—both a paper theory and a computational model. The theory advances common-sense psychology in a number of areas. First, it addresses one of the principal gaps in earlier theories of common-sense psychology; namely, what it is that distinguishes when people ascribe a mind to something from when they don't. Secondly, the theory integrates other related psychological phenomena, notably anthropomorphism, which were previously outside the classical theories of common-sense psychology. Third, the theory focuses on this particular phenomenon of anthropomorphism, and presents what is perhaps the first substantive model of it. And fourth, and

perhaps most importantly for psychology, the complete model of anthropomorphism in common-sense psychology provides some useful theoretical and methodological foundations for further research in cognitive science.

Perhaps the most useful theoretical contribution made to psychology by this thesis is a far more detailed sketch of how different parts need to connect together to form a complete human common-sense psychology. Developing the computational model in chapter 9 showed up a number of weaknesses in existing accounts, namely in the connection between physical and psychological reasoning, in the interactions between anthropomorphism and the intentional stance, and in the interactions between the different factors that make up anthropomorphism. Some tentative suggestions have been made on these points, but this is an area where a lot of work remains to be done. Complete models of the psychology of common-sense psychology are still in their infancy.

Finally, although the thesis never took consciousness as a central concern, there are some important implications for the study of consciousness that come out of the theories I have developed. In particular, I have argued that consciousness is not principally an absolute property, but is ascribed by virtue of people's common-sense psychology. Ascribed consciousness, therefore, is not only observer-relative, but is also subject to all the anthropomorphic factors discussed in chapters 7 and 8. Whatever consciousness really is, the ascription of consciousness closely follows the patterns of common-sense psychology, so the insight into these patterns of ascription offered by the theory and model in this thesis is important to a proper scientific understanding of consciousness. To study consciousness, we need to study how we perceive consciousness in others. Again, this is an area where future work may be especially helpful to the psychology of consciousness.

Implications for artificial intelligence

The implications for artificial intelligence depend on the kind of artificial intelligence that is involved. For practical 'engineering' artificial intelligence, a better understanding of the principles involved in how people actually ascribe mentality to non-human systems will be an important contributor to the design of interfaces to these systems. There is, for example, currently an increase in interest in 'agent' interfaces in human-computer interaction, some of which are directly

anthropomorphic (e.g. Oren, Salomon, Kreitman, & Don, 1990; Walker, Sproull, & Subramani, 1994). A proper understanding of anthropomorphism is important to be able to properly assess the patterns of interaction between people and these agents.

But perhaps more significantly, the theory in this thesis has deep implications for theoretical artificial intelligence. For example, the importance of the observer's common-sense psychology in the ascription of intelligence led to the critique of the alien intelligence hypothesis in chapters 6 and 13. The alien intelligence hypothesis is still, implicitly, a central tenet of artificial intelligence and much cognitive science (McCarthy, 1983; Newell, 1990; Norman, 1981). If, as is suggested by this thesis, the alien intelligence hypothesis is false, the "intelligence in general" (French, 1990) approach to artificial intelligence is bound to fail. Instead, I have argued that for behaviour to be classed as intelligent, it must be what people recognise as intelligent, so it must always be human intelligent behaviour, in some sense at least. This has significant implications. For example, accepting intelligence as something that is implicitly observer-relative means that we need to revise the conventional interpretation of the Turing test, as I have done in chapters 5 and 13. From this different perspective, then, the theory backs up the criticisms of Dreyfus and Dreyfus (1988) and McDermott (1987).

There are other implications for theoretical artificial intelligence. The recent focus on agent systems in distributed artificial intelligence is strongly structured around the sentential forms of beliefs, desires, and intentions (Rao & Georgeff, 1995). The model shows that while it might be possible to describe an agent superficially using sentential beliefs, desires, and intentions, there quickly comes a point where these descriptions have to become so detailed that they no longer bear any real similarity to the kinds of things that we ascribe to people normally in common-sense psychology. Perhaps the most fundamental example of this is the lack of any clear distinction between agents and objects, but there is also a problem with absence of a dispositional element to the conventional belief, desire, and intentional description (Ryle, 1949; Sloman, 1993).

This might seem like rather a heavy price to pay for common-sense psychology, after all, artificial intelligence has managed quite well without doing much in this area for a good long while. Instead, I want to suggest that this fundamental inversion—changing from looking at systems with

smart behaviour to looking at the behaviour which people recognise as smart—could be the key to the problem of common-sense which has dogged artificial intelligence for so long (Dreyfus & Dreyfus, 1988; McDermott, 1987). It stands to reason that before you can build something, you first need to know what to build, and perhaps the central problem of artificial intelligence is that nobody really has any idea what intelligence really is. The solution to this problem is not to *define* it, as some see the Turing test doing (e.g. Millar, 1973), but to *study* it, and it is here that the Turing test really can help. We need to put human common-sense back into artificial intelligence.

On the importance of common-sense psychology

The arguments in this thesis have shown how pervasive common-sense psychology can be; it manifests itself in everything from developmental psychology to the Turing test, from the methodology of thought experiments to the nature of consciousness. All of these are big and nasty problems, and I do not pretend that common-sense psychology is a solution to any, let alone all of them. Instead, I want to suggest that there is a common problem lying behind these, and other, scientific problems—the demon of anthropocentricity. As I argued in chapter 14, there are serious methodological issues associated with the fact that all scientists are human. In some disciplines, such as physics, this may not be such a problem, because there is a substantial gap between the observer and the observed phenomena. Even so, in some branches of physics (some interpretations of quantum physics, for example) the observer is beginning to be seen as having an effect on the understanding of physical phenomena, and as the observer's importance increases, so does the problem of anthropocentricity.

Unfortunately, this same problem is magnified enormously in sciences of the mind. Here the observer and the observed have large parts of their behavioural repertoire in common, and it can become increasingly hard for a scientist to distinguish between ascribed and real unobservables—such as mental states, for example—in another person's behaviour. Methodologically, therefore, it is very important to study what it is in people, especially as scientists, which makes them ascribe

mental states to one another, to animals and to objects like computers and psychological models. Without this, it is going to become increasingly hard to separate real mentality in others from imaginary ascribed mentality in us as observers.

We might consider this by analogy with a microscope. A microscope uses light to study small objects, but there are limits to its resolution—limits imposed not just by the quality of the optics but by the physics of light, by the wavelength of the light used. There comes a point where things that ought in principle to be visible simply aren't. An understanding of the physics of light led to the idea that a different kind of 'light'—electron beams—could overcome this shortcoming. The moral is this: we need to study the medium through which we study things, apart from the study itself, because without this we may find some of the objects of our interest seeming simply to vanish. Using this understanding of the medium, we can change it so that we can study things which we never before thought possible.

All this, by the way, is independent of the eliminative materialist (Churchland, 1981; Stich, 1983) criticism that we should simply forget about common-sense psychology, because common sense has a very different focus. As it happens, I believe that eliminative materialism is false, as do Clark (1989), Searle (1992), and Dennett (1987), and I think that, in particular, Clark's arguments against it are strong. But even if we wanted to accept eliminative materialism, the methodological problems of anthropocentricity and common-sense psychology would not go away. We would *still* need to study common-sense psychology, if only to ensure that we could eliminate it properly.

Common-sense psychology, then, is methodologically important as well as theoretically important. Without a proper understanding of common-sense psychology, we'll increasingly become trapped into the false positives and false negatives of ascribed mentality that have already been seen in artificial intelligence and in philosophical thought experiments, and which form some of the most common methodological problems in this area. But with a proper understanding of these issues, we can begin to become mindful of the traps that our common-sense psychology is setting for us. Common-sense psychology may cause the problem, but understanding it may help us to find the solution.

Anthropomorphism

Anthropomorphism has been studied surprisingly little! When it is mentioned in science, it is usually to be condemned (Kennedy, 1992; Searle, 1992). When it is actually defended (e.g. Caporael, 1986; McCarthy, 1983) it is, however, clearly put within the realm of common-sense psychology, and because this makes it part of a natural faculty, there doesn't seem to be any way that we can switch it off. The problem of anthropomorphism seems to be unavoidable (Krementsov & Todes, 1991). So, building on the experimental work of Tamir and Zohar (1991) and, especially, Eddy *et al.* (1993), along with the theoretical contributions of Caporael (1986), I have suggested that anthropomorphism actually plays the role of the rationality assumption in Dennett's model of intentional systems (Dennett, 1971)—despite the fact that anthropomorphism has nothing to do with rationality. Playing the role of the rationality assumption, it is anthropomorphism that governs when the intentional stance is taken—and therefore when mentality is ascribed—to a system.

The model of anthropomorphism in this thesis consists of seven dispositional factors; depending on the similarity and the familiarity of the system to the ascriber, the animacy of the system, the ascriber's beliefs about the system's structure, the interaction medium, the context, and the reason why the ascriber is taking a stance. These factors are not 'absolute' in any sense; that is, they depend on a kind of compatibility between the system and the ascriber, rather than on special properties of the system. In this, they correspond more to Dawkins' (1989) "armpit effect" altruism than his "green-beard effect" altruism. This leads to a new view of the way a system's properties affect how people ascribe mentality to it—a 'lock and key' metaphor which I'll discuss in the next section.

Identifying anthropomorphism with Dennett's (1971) rationality assumption goes some way to defusing those criticisms of his theory (e.g. Baker, 1994; Fodor, 1985) which point out that the rationality assumption had nothing to do with rationality. Of course it doesn't, but Dennett never claimed that it really did refer to rationality in anything more than a metaphorical "pre-theoretical" sense. Building on this identification, I have appropriated Dennett's (1971) philosophical model of intentional systems, along with this model of anthropomorphism, into a model of common-sense psychology which has one significant advantage: as well as describing how mental states can be ascribed

to others, it also describes how to decide *when* mental states will be ascribed in the first place. This model, in the computational form described in chapter 9, forms the core of the models for the false belief test and Searle's Chinese Room, developed and discussed in the third part of this thesis.

The 'lock and key' metaphor

The thesis' revised interpretation of the Turing test as evaluating the compatibility between the system's behaviour and the observer's common-sense psychology has one very significant implication. Instead of 'intelligence' or 'consciousness' being absolute phenomena, they are, in effect, a measure of relative psychological compatibility. We can imagine this through the metaphor of a lock and a key. Imagine that there are in the world, many locked doors and many keys, all of which are different, some only very slightly and some radically. If we try to open a door with a key that is very different from the right key for its lock, we fail, and the door remains closed to us. If we try to open a door with the right key, we can open the door easily. But there will be some keys that differ just enough to make opening that door possible, but a bit harder; we might need to wiggle the key a bit before we can open the door.

This is, I believe, a good metaphor for intelligence and consciousness. Different people, animals, computers, and so on, behave in ways that correspond to doors. Every individual person, similarly, has a common-sense psychology, which corresponds to a key. If the key opens the door—if the common-sense psychology can 'read' the behaviour—the person ascribes mentality to the system. But it is important to remember that it is only *people* that have keys, so there will be many doors that always remain closed; these correspond to the alien intelligences that I mentioned in chapter 6. According to this model, alien intelligence, and even alien consciousness, are entirely possible and may even exist already in computers today. This is, however, mostly a philosophical irrelevance, as it is only those systems whose behaviours that can be psychologically read by us humans that will, in practice, be properly seen as having intelligence or consciousness. If artificial intelligence is to be seen as real intelligence or consciousness, it must provide a door which can be opened by many people's keys—behaviours which can be read by human common-sense psychology. It must become human intelligence.

The implications of this 'lock and key' metaphor are to firmly reconnect artificial intelligence with human psychology—and especially with human common-sense psychology—and to banish the ideas of "intelligence in general" (French, 1990) and alien intelligence as worthwhile goals in their own right, to the bin of irrelevant myths, along with the perpetual motion machine. Furthermore, it focuses on the problem of consciousness in particular; not only must science account for consciousness in general—it must also ground that with an account of human consciousness in particular, and how human consciousness manifests itself in human behaviour; without that, it will never be possible to account for how people recognise consciousness even when it does exist.

As I've said before, I do not want to claim that all properties are observer-relative, only that common-sense psychological ones are. This is inevitable, given that common-sense psychology is the window through which one person thinks about another. And as I've shown, common-sense psychology is endemic to our concepts of 'intelligence' and 'consciousness', so in great measure these must be observer-relative. These concepts are only truly meaningful within the frame of reference of our common-sense psychology; they *cannot* be objective absolutes.

The false belief test

I do not want to suggest that the model in this thesis contributes new theories to common-sense psychology in the false belief test. Indeed, research in this field is so active that new kinds of false belief test appear almost by the month. I do believe, however, that it offers a new methodology to the field; a methodology which can help to compare and contrast different theories as they are proposed. In this thesis, I have only compared four theories of the actual structure of common-sense psychology, Leslie's (1987) theory of mind mechanism, Gordon's (1986) simulation theory, Chandler's (Chandler & Boyes, 1982) copy theory, and Perner's (1991) situation theory. Of these, the model that comes best out of the comparison—within the limited scope of this modelling framework—seems to be Perner's situation theory. This is because it shows best the character of 'theory extension'—in the situation theory the transition from a model which fails the false belief test to a model that passes it is relatively small and self-contained.

An improved modelling framework, especially one including desires and intentions, would allow more models to be compared more effectively. But even as it stands, this framework makes comparisons between the different theories easier to make, by offering an additional discriminatory tool over and above philosophical argument and psychological experiment. Although the technique of computational modelling cannot be counted on to produce hard and fast models in the absence of empirical results, it can complement experimental psychology, offering new hypotheses and an additional format for criticising models (as in chapter 10).

The Turing test and the Chinese Room

In this thesis, I have not even attempted to defend the Turing test as an operational test for intelligence; the operational interpretation, and even Dennett's (1985) quasi-legal, interpretation, do create a tendency for artificial intelligence to build purely imitative systems, a methodology that is neither scientifically useful nor a good advertisement for the discipline (Hayes & Ford, 1995). Moor's (1976) inductive interpretation is different in this respect, it means we can continue to use the test as a methodological tool, to look for those behaviours that increase the likelihood that we see intelligence, without emphasising the mechanisms that generate those behaviours. This is in line with the 'lock and key' metaphor; instead of investigating the abstract properties of locks on their own, it also looks at which keys tend to open which locks. In fact, practical use of the Turing test in artificial intelligence—trying to build locks and seeing which keys fit them—is clearly complementary to studying only the existing locks.

The Chinese Room is a good example of this. In this thesis, I have made a systematic analysis of the effects of different people's intuitions on the thought experiment aspect of Searle's Chinese Room argument. This is consistent with the methodology I am advocating with respect to the Turing test, and even this one intuition pump, with only a few variations on the original format of the thought experiment, has emphasised and clarified several important parts of the model of common-sense psychology developed in this thesis. That is, by carrying out a kind of philosophical Turing test, some of the important correlations between different people's keys and locks have become substantially clearer. The Turing test, and the Chinese Room, are not worthless debates

to be discarded and ignored (Hayes & Ford, 1995), they offer a new tool for studying common-sense psychology and psychology in general; a tool which perhaps should be developed and enhanced in experimental psychology to make it even more useful in this area.

The inverted Turing test proposed in chapter 13 is a more philosophical version of this position. It inverts the Turing test, putting systems in the position of the judge analysing human behaviour, rather than that of the player imitating it. Underneath the inverted Turing test is the position that building models of the processes involved in the human ascription of mentality may help us to understand the nature of what we call intelligence and of what we call consciousness. In this sense, the inverted Turing test rationalises, and sets the general context for, the models in the third part of this thesis. It is, of course, not intended to be a practical approach to the development of intelligent systems; it is an inductive tool just like the original Turing test. Its role is this: just as the Turing test can be thought of as a reference comparator for general intelligent behaviour, evaluated linguistically, the inverse Turing test can be thought of as a specific reference comparator for the behaviours of common-sense psychology.

Similarly, a more systematic and complete review of the different replies and debates on the Chinese Room may continue to offer new insights into the role of common-sense psychology, not only in this particular thought experiment, but in other areas involving the ascription of mental states. On this basis, unlike many (e.g. Bringsjord, 1995; French, 1990; Hayes & Ford, 1995; Whitby, 1996) I do not want to abandon the Turing test. Indeed, I think that only from thought experiments like those of Turing and Searle can we really learn how different architectures affect our ascription of mentality in practice, as they certainly seem to do. The fact that the Turing test's apparent behaviourism runs counter to our intuitions tells us about our intuitions as well as about the Turing test, and knowledge about these intuitions is knowledge that cognitive science badly needs.

There is much to be learnt from more detailed empirical study of people's recognition of intelligent behaviour, without simply assuming intelligent behaviour to be a universal absolute. The Turing test is one very sharp tool that we can use to study people's intuitions under carefully controlled conditions. We can look at these phenomena, but without necessarily committing ourselves to the Turing test format, or anything like it. Essentially, I would argue that people are

fundamentally anthropocentric, and that one way—perhaps the only way—of overcoming this is to study this anthropocentricity in its own right. Through this study, we can, in principle, discover the effects of that anthropocentricity and then, in Caporael's terms, "set traps for it" (Caporael, 1986). Furthermore, if we fail to study this anthropocentricity, it may continue to set traps for us.

Models and architectures

One of the more significant results of this thesis is that some apparently scientific concepts of cognitive science can emerge from common-sense psychology. One of the most obvious of these is the strong connection between levels of explanation, in the manner of Newell (1992), and the structure component in the model of anthropomorphism. This implies that some of the more important concepts of the information processing model, in particular, such as the virtual machine, may live as much in the head of the ascriber as they do in any real behaviours of the system.

The information processing metaphor is the most obvious victim to this problem, but it is not the only one—and anyway, it is a bit of an easy target these days. In chapter 14, I've described three general "normalising" (Turkle, 1988) strategies for psychological models. First, there is the 'similarity fallacy' which suggests that things tend to be more like us mentally if they are more like us physically. Second, there is the 'consciousness box fallacy' which provides some evidence for the intuitive plausibility of homunculi in the information processing model. And third, there is the 'radical emergence fallacy' which shows the psychological attractiveness of emergent models, both at the neurophysiological level and at the quantum level. By recognising these strategies, it becomes clearer that the main difference between information processing models and emergent models is not so much to do with their actual implementations, but to do with the appeal of their implementations to our common-sense psychology. That is, functionally different architectures may behave identically but still be seen differently. This could be seen as an argument for functionalism, but unfortunately it isn't, because the flip-side of this phenomenon is that functionally identical architectures can also be seen differently! And to cap it all, individual differences between people's perceptions of different architectures also play a significant part in how they are seen in practice.

These results might seem somewhat negative, but I believe that it is important to state the problem clearly. But there is a positive side; as well as causing the problem, common-sense psychology may help us find the solution. A systematic analysis of people's intuitions regarding these different models and architectures can at least help us to become aware of the intuitive responses that people are likely to have to our models—and we can use this to ensure that these responses don't get in the way of the models themselves. In effect, we can take Caporael's (1986) strategy of "setting traps" for common-sense psychology, and apply it to the design of our models, metaphors, and architectures.

Summary

So we have come quite a roundabout journey in this look at common-sense psychology. I have argued that it is both methodologically and theoretically important, and I have advanced the field a little at an important junction between the two—anthropomorphism. Anthropomorphism certainly seems to be an important phenomenon; methodologically, it deeply affects our intuitions in thought experiments and cognitive models, and theoretically, it plays an essential role in common-sense psychology, in that it seems to be the key to our distinction between psychological and physical objects. All in all, perhaps, more questions have come out of this study than have answers, but I believe that this is still an advance, in that these questions might not have been recognised as either relevant or important to cognitive science before this theory or model was developed.

If anybody wants a message to take away from this thesis, I would like it to be this general feeling of the importance of common-sense psychology. Common-sense psychology is too important to let anybody eliminate it from science. It is active in our perception of each other, as people, but it is also active in our perception, *as scientists*, of psychological models, implementation architectures, and thought experiments. Common-sense psychology is a natural, endemic, and inescapable part of how we see the world, each other, and ourselves.

BLANK IN ORIGINAL

Bibliography

Akins, K. A. (1993). A Bat Without Qualities? In M. Davies & G. W. Humphreys (Eds.), *Consciousness* (pp. 258-273). Oxford: Blackwell.

Amit, D. J. (1989). *Modelling Brain Function: The World of Attractor Neural Networks*. Cambridge: Cambridge University Press.

Anderson, J. R. (1993). *Rules of the Mind*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Ashmore, M. (1993). *Behaviour Modification of a Catflap*. Paper presented at the workshop Non-human Agency: A Contradiction in Terms, University of Surrey, September 1993.

Astington, J. W., & Gopnik, A. (1991). Theoretical Explanations of Children's Understanding of the Mind. *British Journal of Developmental Psychology, 9*, 7-31.

Austin, J. L. (1962). *How To Do Things With Words*. Cambridge, Massachusetts: Harvard University Press.

Baddeley, A. D., & Hitch, G. (1974). Working Memory. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (Vol. 8). London: Academic Press.

Bailey, J. (1992). First We Reshape Our Computers, Then Our Computers Reshape Us: The Broader Intellectual Impact of Parallelism. *Daedalus, 121*(1), 67-86.

Baker, L. R. (1994). Instrumental Intentionality. In S. P. Stich & T. A. Warfield (Eds.), *Mental Representation* (pp. 332-344). Oxford: Basil Blackwell.

Baron-Cohen, S. (1989). The Autistic Child's Theory of Mind: A Case of Specific Developmental Delay. *Journal of Child Psychology and Psychiatry, 30*, 285-297.

Baron-Cohen, S. (1993). From Attention-Goal Psychology to Belief-Desire Psychology: the Development of a Theory of Mind and its Dysfunction. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen, J. (Eds.), *Understanding Other Minds: Perspectives From Autism* (pp. 59-82). Oxford: Oxford University Press.

Baron-Cohen, S. (1995). *Mindblindness*. Cambridge, Massachusetts: MIT Press.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the Autistic Child have a 'Theory of Mind'? *Cognition, 21*(1), 37-46.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1986). Mechanical, Behavioural, and Intentional Understanding of Picture Stories in Autistic Children. *British Journal of Developmental Psychology, 4*, 113-125.

Bechtel, W. (1992). Studying the Thinking of Non-Human Animals. *Biology and Philosophy, 7*, 209-215.

Bennett, J. (1978). Some Remarks About Concepts. *Behavioral and Brain Sciences, 1*, 557-560.

Black, M. (1979). More About Metaphor. In A. Ortony (Ed.), *Metaphor and Thought* (pp. 19-43). Cambridge: Cambridge University Press.

Block, N. (1978). Troubles With Functionalism. In C. W. Savage (Ed.), *Perception and Cognition: Issues in the Foundations of Psychology*. Minneapolis: University of Minnesota Press.

Block, N. (1980). What Intuitions About Homunculi Don't Show. *Behavioral and Brain Sciences, 3*, 425-426.

Block, N. (1981). Psychologism and Behaviourism. *Philosophical Review, 40*, 5-43.

Block, N. (1986). Advertisement for a Semantics for Psychology. *Midwest Studies in Philosophy, 10*, 615-678.

Boden, M. A. (1988). *Computer Models of Mind*. Cambridge: Cambridge University Press.

Bohm, D. (1980). *Wholeness and the Implicate Order*. Routledge.

Boster, J., Berlin, B., & O'Neill, J. (1986). The Correspondence of Jivaroan to Scientific Ornithology. *American Anthropologist, 88*, 569-583.

Bowlby, J. (1969). *Attachment and Loss, Volume I: Attachment*. New York: Hogarth Press.

Boyce, A. J. (1969). *Mapping Diversity*. Paper presented at the Colloquium in Numerical Taxonomy, University of St. Andrews.

Boyd, R. (1979). Metaphor and Theory Change. In A. Ortony (Ed.), *Metaphor and Thought* (pp. 356-408). Cambridge: Cambridge University Press.

Bringsjord, S. (1995). Could, How Could We Tell if, and Why Should—Androids Have Inner Lives? In K. M. Ford, C. Glymour, & P. J. Hayes (Eds.), *Android Epistemology* (pp. 93-122). Cambridge, Massachusetts: AAAI Press/MIT Press.

Brooks, R. A. (1991a). Elephants Don't Play Chess. In P. Maes (Ed.), *Designing Autonomous Agents: Theory and Practice from Biology to Engineering and Back* (pp. 3-15). Cambridge, Massachusetts: MIT Press.

Brooks, R. A. (1991b). *Intelligence Without Reason*. Paper presented at the Twelfth International Joint Conference on Artificial Intelligence (IJCAI'91).

Brooks, R. A., & Stein, L. A. (1993). *Building Brains for Bodies* (AI Memo 1439): MIT AI Laboratory.

Bruner, J., & Feldman, C. (1993). Theories of Mind and the Problems of Autism. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen, J. (Eds.), *Understanding Other Minds: Perspectives from Autism* (pp. 267-291). Oxford: Oxford University Press.

Byrne, R. W., & Whiten, A. (Eds.). (1988). *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans.* Oxford: Oxford University Press.

Calendra, A. (1964). The Barometer Story. *Current Science, 49*(14), 6-10.

Caporael, L. R. (1986). Anthropomorphism and Mechanomorphism: Two Faces of the Human Machine. *Computers in Human Behavior, 2*(3), 215-234.

Carey, S. (1985). *Conceptual Change in Childhood.* Cambridge, Massachusetts: MIT Press.

Casey, G. (1992). Minds and Machines. *American Catholic Philosophical Quarterly, LXVI*(1), 57-80.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Chandler, M. J., & Boyes, M. (1982). Social-Cognitive Development. In B. B. Wolman (Ed.), *Handbook of Developmental Psychology* (pp. 387-402). Englewood Cliffs, New Jersey: Prentice-Hall.

Cheney, D., & Seyfarth, R. M. (1990). *How Monkeys See the World.* Chicago: University of Chicago Press.

Cheney, D., & Seyfarth, R. M. (1991). Reading Minds or Reading Behaviour: Tests for a Theory of Mind in Monkeys. In A. Whiten (Ed.), *Natural Theories of Mind: Evolution, Development, and Simulation of Everyday Mindreading* (pp. 175-194). Oxford: Basil Blackwell.

Churchland, P. A. (1990). Could a Machine Think? *Scientific American, 262*, 32-37.

Churchland, P. M. (1979). *Scientific Realism and the Plasticity of Mind.* Cambridge: Cambridge University Press.

Churchland, P. M. (1981). Eliminative Materialism and the Propositional Attitudes. *Journal of Philosophy, 78*, 67-90.

Clark, A. (1987). From Folk Psychology to Naive Psychology. *Cognitive Science, 11*, 139-154.

Clark, A. (1988). *Two Kinds of Cognitive Science?* Paper presented at the European Conference on Artificial Intelligence, ECAI'88.

Clark, A. (1989). *Microcognition.* Cambridge, Massachusetts: MIT Press.

Cohen, P. R., & Levesque, H. J. (1987). *Persistence, Intention and Commitment* (CSLI-87-88): Center for the Study of Language and Information.

Cohen, P. R., & Levesque, H. J. (1990). Intention is Choice with Commitment. *Artificial Intelligence, 42*, 213-261.

Colby, K. M. (1981). Modeling a Paranoid Mind. *Behavioural and Brain Sciences, 4*, 515-560.

Collins, H. M. (1990). *Artificial Experts: Social Knowledge and Intelligent Machines.* Cambridge, Massachusetts: MIT Press.

Darwin, C. (1871). *The Descent of Man.* London: John Murray.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals.* London: John Murray.

Davies, M. (1994). The Mental Simulation Debate. In C. Peacocke (Ed.), *Objectivity, Simulation and the Unity of Consciousness* (pp. 99-128). Oxford: Oxford University Press for the British Academy.

Davis, E. (1988). *Inferring Ignorance from the Locality of Visual Perception.* Paper presented at the Seventh National Conference on Artificial Intelligence, AAAI'88.

Davis, E. (1992). Commonsense Reasoning. In S. C. Shapiro (Ed.), *Encyclopaedia of Artificial Intelligence* (pp. 1288-1294). New York: John Wiley and Sons.

Dawkins, R. (1982). *The Extended Phenotype.* Oxford: Oxford University Press.

Dawkins, R. (1989). *The Selfish Gene.* Oxford: Oxford University Press.

Dennett, D. C. (1971). Intentional Systems. *Journal of Philosophy, 68*, 87-106.

Dennett, D. C. (1978). Beliefs About Beliefs. *Behavioral and Brain Sciences, 1*, 568-570.

Dennett, D. C. (1980). The Milk of Human Intentionaity. *Behavioral and Brain Sciences, 3*, 428-430.

Dennett, D. C. (1982). Beyond Belief. In A. Woodfield (Ed.), *Thought and Object.* Oxford: Clarendon Press.

Dennett, D. C. (1984). Cognitive Wheels: The Frame Problem of AI. In C. Hookway (Ed.), *Minds, machines, and evolution* (pp. 129-151). Cambridge: Cambridge University Press.

Dennett, D. C. (1985). Can Machines Think? In M. Shafto (Ed.), *How We Know* (pp. 121-145). New York: Harper and Row.

Dennett, D. C. (1987). *The Intentional Stance.* Cambridge, Massachusetts: MIT Press.

Dennett, D. C. (1991). *Consciousness Explained.* Little, Brown, and Company.

Dreyfus, H. L., & Dreyfus, S. E. (1988). Making a Mind Versus Modelling the Brain: Artificial Intelligence Back at a Branchpoint. *Daedalus, 117*(1), 185-197.

Eccles, J. C. (1964). Cited in R. W. Sperry, Consciousness and Causality. In R. L. Gregory (Ed.), *The Oxford Companion to the Mind.* Oxford: Oxford University Press.

Eddy, T. J., Gallup, G. G., & Povinelli, D. J. (1993). Attribution of Cognitive States to Animals: Anthropomorphism in Comparitive Perspective. *Journal of Social Issues, 49*(1), 87-101.

Elias, N. (1974). The Sciences: Towards a Theory. In R. Whitley (Ed.), *Social Processes of Scientific Development* (pp. 21-42). London: Routledge and Kegan Paul.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development.* Cambridge, Massachusetts: MIT Press.

Elton, M. (1993). *Human and Animal Consciousness* (Cognitive Science Research Paper 287): School of Cognitive and Computing Sciences, University of Sussex.

Federn, P. (1952). *Ego Psychology and the Psychoses.* New York: Basic Books.

Flanagan, O. (1993). *Consciousness Reconsidered.* Cambridge, Massachusetts: MIT Press.

Fodor, J. A. (1980). Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology. *Behavioural and Brain Sciences, 3*, 63-73.

Fodor, J. A. (1981). *Representations.* Cambridge, Massachusetts: MIT Press.

Fodor, J. A. (1985). Fodor's Guide to Mental Representation: The Intelligent Auntie's Vade-Mecum. *Mind, 94*, 55-97.

Fodor, J. A. (1986). Why Paramecia Don't Have Mental Representations. *Midwest Studies in Philosophy, 10*, 3-23.

Fodor, J. A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind.* Cambridge, Massachusetts: MIT Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition, 20*, 1988.

French, R. M. (1990). Subcognition and the Limits of the Turing Test. *Mind, 99*, 53-65.

Garfinkel, H. (1967). *Studies in Ethnomethodology.* Englewood Cliffs, New Jersey: Prentice-Hall.

Genesereth, M. R., & Nilsson, N. J. (1987). *Logical Foundations of Artificial Intelligence.* Los Altos, California: Morgan Kaufmann.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995) Taking the Intentional Stance at 12 Months of Age. *Cognition, 56,* 165-193.

Goldman, A. I. (1993). The Psychology of Folk Psychology. *Behavioural and Brain Sciences, 16*, 15-28.

Gómez, J. C. (1991). Visual Behaviour as a Window for Reading the Mind of Others in Primates. In A. Whiten (Ed.), *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading* (pp. 195-208). Oxford: Oxford University Press.

Gómez, J. C., Sarriá, E., & Tamarit, J. (1993). The Comparative Study of Early Communication and Theories of Mind: Ontogeny, Phylogeny, and Pathology. In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding Other Minds: Perspectives From Autism* (pp. 397-426). Oxford: Oxford University Press.

Gordon, R. M. (1986). Folk Psychology as Simulation. *Mind and Language, 1*(2), 158-171.

Graham, G. (1987). The Origins of Folk Psychology. *Inquiry, 30*, 357-379.

Grice, H. P. (1957). Meaning. *Philosophical Review, 66*, 377-388.

Gunderson, K. (1971). *Machines and Mentality*. New York: Doubleday.

Haas, A. R. (1986). A Syntactic Theory of Belief and Knowledge. *Artificial Intelligence, 23*(3), 242-292.

Hamilton, W. D. (1964). The Genetical Evolution of Social Behaviour II. *Journal of Theoretical Biology, 7*, 17-32.

Haraway, D. (1992). Otherworldly Conversations; Terran Topics; Local Terms. *Science as Culture, 3*(14), 64-98.

Harman, G. (1978). Studying the Chimpanzee's Theory of Mind. *Behavioural and Brain Sciences, 1*, 591.

Harnad, S. (1990). The Symbol Grounding Problem. *Physica D, 42*, 335-346.

Harnad, S. (1991). Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem. *Minds and Machines, 1*(1), 43-54.

Harnad, S. (1992). The Turing Test is not a Trick: Turing Indistinguishability is a Scientific Criterion. *SIGART Bulletin, 3*(4), 9-10.

Harris, P. (1993). Pretending and Planning. In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding Other Minds: Perspectives From Autism* (pp. 228-246). Oxford: Oxford University Press.

Harris, P. L. (1989). *Children and Emotion: The Development of Psychological Understanding*. Oxford: Basil Blackwell.

Harris, P. L., Brown, E., Marriot, C., Whittall, S., & Harmer, S. (1991). Monsters, Ghosts, and Witches: Testing the Limits of the Fantasy—Reality Distinction. *British Journal of Developmental Psychology, 9*, 105-124.

Haugeland, J. (1978). The Nature and Plausibility of Cognitivism. *Behavioural and Brain Sciences, 1*, 215-226.

Haugeland, J. (1980). Programs, Causal Powers, and Intentionality. *Behavioral and Brain Sciences, 3*, 432-433.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. Cambridge, Massachusetts: MIT Press.

Hauser, L. (1993). Reaping the Whirlwind: Reply to Harnad's 'Other Bodies, Other Minds'. *Minds and Machines, 3*(2), 219-237.

Hayes, P. J. (1985a). Naive Physics I: Ontology for Liquids. In J. R. Hobbs & R. C. Moore (Eds.), *Formal Theories of the Commonsense World* (pp. 71-108). Norwood, New Jersey: Ablex.

Hayes, P. J. (1985b). The Second Naive Physics Manifesto. In J. R. Hobbs & R. C. Moore (Eds.), *Formal Theories of the Commonsense World* (pp. 1-36). Norwood, New Jersey: Ablex.

Hayes, P. J., & Ford, K. M. (1995). *Turing Test Considered Harmful* (pp. 972-977). In the proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI'95), Morgan-Kaufmann.

Hayes, P. J., Ford, K. M., & Agnew, N. (1994). On Babies and Bathwater—A Cautionary Tale. *AI Magazine, 15*(4), 15-26.

Heal, J. (1994). Simulation vs. Theory Theory: What is at Issue? In C. Peacocke (Ed.), *Objectivity, Simulation and the Unity of Consciousness* (pp. 129-144). Oxford: Oxford University Press for the British Academy.

Heider, F. (1958). *The Psychology of Interpersonal Relations.* New York: Wiley.

Hinde, R. A. (1974). *Biological Bases of Human Social Behaviour.* New York: McGraw-Hill.

Hobbs, J. R., & Moore, R. C. (Eds.). (1985). *Formal Theories of the Commonsense World.* Norwood, New Jersey: Ablex.

Hobson, P. (1993). Understanding Persons: The Role of Affect. In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding Other Minds: Perspectives From Autism* (pp. 204-227). Oxford: Oxford University Press.

Hofstadter, D. R. (1985). *Metamagical Themas: Questing for the Eessence of Mind and Pattern.* New York: Basic Books.

Hofstadter, D. R., & Dennett, D. C. (1981). *The Mind's I: Fantasies and Reflections on Self and Soul.* New York: Basic Books.

Humphrey, N. K. (1976). The Social Function of Intellect. In P. P. G. Bateson & R. A. Hinde (Eds.), *Growing Points in Ethology* (pp. 303-317). Cambridge: Cambridge University Press.

Humphrey, N. K. (1984). Having Feelings and Showing Feelings, *Consciousness Regained* (pp. 42-45). Oxford: Oxford University Press.

Humphrey, N. K. (1992). *A History of the Mind.* London: Chatto and Windus.

Inagaki, K., & Hatano, G. (1991) Constrained Person Analogy in Young Children's Biological Inference. *Cognitive Development, 6,* 219-231.

Israel, D. (1985). A Short Companion to the Naive Physics Manifesto. In J. R. Hobbs & R. C. Moore (Eds.), *Formal Theories of the Commonsense World* (pp. 427-447). Norwood, New Jersey: Ablex.

Jackendoff, R. (1985). Information is in the Mind of the Beholder. *Linguistics and Philosophy, 8,* 23-33.

Johnson, C. N. (1988). Theory of Mind and the Structure of Conscious Experience. In J. W. Astington, P. L. Harris, & D. R. Olson (Eds.), *Developing Theories of Mind.* Cambridge: Cambridge University Press.

Johnson, M. (1987). *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Cognition.* Chicago: University of Chicago Press.

Kaplan, D. (1980). *Demonstratives.* The John Locke Lectures, Oxford University.

Kay, A. (1990). User Interface: A Personal View. In B. Laurel & S. J. Mountford (Eds.), *The Art of Human-Computer Interface Design* (pp. 191-207). Reading, Massachusetts: Addison-Wesley.

Kennedy, J. S. (1992). *The New Anthropomorphism.* Cambridge: Cambridge University Press.

Kolata, G. (1982). How Can Computers Get Common Sense? *Science, 217,* 1237-1238.

Krementsov, N. L., & Todes, D. P. (1991). On Metaphors, Animals, and Us. *Journal of Social Issues, 47*(3), 67-81.

Kühberger, A., Perner, J., Schulte, M., & Leingruber, R. (1995). Choice or No Choice: Is the Langer Effect Evidence Against Simulation? *Mind and Language, 10*(4), 423-436.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things.* Chicago: University of Chicago Press.

Langer, E. J. (1975). The Illusion of Control. *Journal of Personality and Social Psychology, 32*(2), 311-328.

Lenat, D. B., Prakash, M., & Shepherd, M. (1986). CYC: Using Commonsense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks. *AI Magazine, 6,* 65-85.

Leslie, A. M. (1987). Pretence and Representation: the Origins of 'Theory of Mind'. *Psychological Review, 94*(412-426).

Leslie, A. M. (1988). Some Implications of Pretense for Mechanisms Underlying the Child's Theory of Mind. In J. W. Astington, P. L. Harris, & D. R. Olson (Eds.), *Developing Theories of Mind* (pp. 19-46). Cambridge: Cambridge University Press.

Leslie, A. M. (1991). The Theory of Mind Impairment in Autism: Evidence for a Modular Mechanism of Development. In A. Whiten (Ed.), *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading* (pp. 63-78). Oxford: Basil Blackwell.

Leslie, A. M. (1995). A Theory of Agency. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal Cognition: A Multidisciplinary Debate* (pp. 121-141). Oxford: Oxford University Press.

Leslie, A. M., & Roth, D. (1993). What Autism Teaches Us About Metarepresentation. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen, J. (Eds.), *Understanding Other Minds: Perspectives from Autism* (pp. 83-111). Oxford: Oxford University Press.

Leslie, A. M., & Thiass, L. (1992). Domain Specificity in Conceptual Development: Neuropsychological Evidence from Autism. *Cognition, 43,* 225-251.

Lord, C. (1993). The Complexity of Social Behaviour in Autism. In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding Other Minds: Perspectives From Autism* (pp. 292-316). Oxford: Oxford University Press.

Lucas, M. M., & Hayes, P. J. (Eds.). (1982). *Proceedings of the Cognitive Curricula Conference.* University of Rochester.

Matthews, R. J. (1984). Troubles with Representationalism. *Social Research, 51*(4), 1065-1097.

McCarthy, J. (1959). Programs with Common Sense, *Proceedings of the Teddington Conference on the Mechanisation of Thought Processes* (pp. 75-91). London: HMSO.

McCarthy, J. (1977). *Epistemological Problems of Artificial Intelligence.* Paper presented at the Fifth International Joint Conference on Artificial Intelligence (IJCAI'77).

McCarthy, J. (1979). Ascribing Mental Qualities to Machines. In M. Ringle (Ed.), *Philosophical Perspectives in Artificial Intelligence* (pp. 161-195). Brighton: Harvester Press.

McCarthy, J. (1983). The Little Thoughts of Thinking Machines. *Psychology Today, 17*(12).

McCarthy, J. (1988). Mathematical Logic in Artificial Intelligence. *Daedalus, 117*(1), 297-311.

McCarthy, J., & Hayes, P. J. (1969). Some Philosophical Problems From the Standpoint of Artificial Intelligence. In B. Meltzer & D. Michie (Eds.), *Machine Intelligence 4* (pp. 463-502). Edinburgh: Edinburgh University Press.

McDermott, D. (1987). A Critique of Pure Reason. *Computational Intelligence, 3*, 151-160.

McGinn, C. (1989). Can We Solve the Mind-Body Problem? *Mind, 98*, 349-366.

McLelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The Appeal of Parallel Distributed Processing. In J. L. McLelland & D. E. Rumelhart (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Vol. 1). Cambridge, Massachusetts: MIT Press.

Meltzoff, A. (1995). Understanding the Intentions of Others: Re-Enactment of Intended Acts by 18-Month-Old Children. *Developmental Psychology, 31*(5), 838-850.

Meltzoff, A., & Gopnik, A. (1993). The Role of Imitation in Understanding Persons and Developing a Theory of Mind. In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding Other Minds: Perspectives from Autism* (pp. 335-366). Oxford: Oxford University Press.

Michie, D. (1993). Turing's Test and Conscious Thought. *Artificial Intelligence, 60*(1).

Millar, P. H. (1973). On the Point of the Imitation Game. *Mind, 82*, 595-597.

Miller, G. (1981). Trends and Debates in Cognitive Psychology. *Cognition, 10*, 215-225.

Miller, G. (1983). Cited in J. Miller (Ed.), *States of Mind.* New York: Pantheon.

Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories.* Cambridge, Massachusetts: MIT Press.

Millikan, R. G. (1993). *White Queen Psychology and Other Essays for Alice.* Cambridge, Massachusetts: MIT Press.

Minsky, M. (1981). A Framework for Representing Knowledge. In J. Haugeland (Ed.), *Mind Design* (pp. 95-128). Cambridge, Massachusetts: MIT Press.

Minsky, M. (1985). *The Society of Mind.* New York: Simon and Schuster.

Moor, J. H. (1976). An Analysis of the Turing Test. *Philosophical Studies, 30,* 249-257.

Moor, J. H. (1992). Turing Test. In S. C. Shapiro (Ed.), *Encyclopaedia of Artificial Intelligence* (pp. 1626-1629). New York: John Wiley and Sons.

Moore, R. C. (1985). A Formal Theory of Knowledge and Action. In J. R. Hobbs & R. C. Moore (Eds.), *Formal Theories of the Commonsense World* (pp. 319-358). Norwood, New Jersey: Ablex.

Nagel, T. (1974). What is it Like to be a Bat? *Philosophical Review, LXXXIII.*

Narayanan, A. (1996). The Intentional Stance and the Imitation Game. In P. Millican & A. Clark (Eds.), *Machines and Thought: The Legacy of Alan Turing* (Vol. 1, pp. 63-79). Oxford: Clarendon Press.

Nass, C., Steuer, J., & Tauber, E. R. (1994). *Computers are Social Actors.* Paper presented at CHI'94.

Newell, A. (1990). *Unified Theories of Cognition.* Cambridge, Massachusetts: Harvard University Press.

Newell, A. (1992). Precis of 'Unified Theories of Cognition'. *Behavioural and Brain Sciences,* 425-492.

Newell, A., & Simon, H. A. (1976). Computer Science as Empirical Enquiry: Symbols and Search. *Communications of the ACM, 19.*

Norman, D. A. (1981). What is Cognitive Science? In D. A. Norman (Ed.), *Perspectives on Cognitive Science.* Norwood, New Jersey: Ablex.

Norman, D. A., & Shallice, T. (1980). *Attention to Action: Willed and Automatic Behaviour* (CHIP Report 99). San Diego, CA: University of California, San Diego.

Oren, T., Salomon, G., Kreitman, K., & Don, A. (1990). Guides: Characterizing the Interface. In B. Laurel & S. J. Mountford (Eds.), *The Art of Human-computer Interface Design* (pp. 367-381). Reading, Massachusetts: Addison-Wesley.

Ortony, A. (1979). The Role of Similarity in Similes and Metaphors. In A. Ortony (Ed.), *Metaphor and Thought.* Cambridge: Cambridge University Press.

Papert, S. (1988). One AI or Many? *Daedalus, 117.*

Penny, S. (1993). *Petit Mal* [Robot]: See URL: *http://www-art.cfa.cmu.edu/www-penny/works/petitmal/petitcode.html.*

Penrose, R. (1989). *The Emperor's New Mind.* Oxford: Oxford University Press.

Penrose, R. (1994). *Shadows of the Mind.* Oxford: Oxford University Press.

Perner, J. (1991). *Understanding the Representational Mind.* Cambridge, Massachusetts: MIT Press.

Perner, J. (1993). The Theory of Mind Deficit in Autism: Rethinking the Metarepresentation Theory. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen, J. (Eds.), *Understanding Other Minds: Perspectives from Autism* (pp. 112-137). Oxford: Oxford University Press.

Perner, J. (1994). The Necessity and Impossibility of Simulation. In C. Peacocke (Ed.), *Objectivity, Simulation and the Unity of Consciousness* (pp. 145-154). Oxford: Oxford University Press for the British Academy.

Premack, D., & Dasser, V. (1991). Theory of Mind in Apes and Children. In A. Whiten (Ed.), *Natural Theories of Mind: Evolution, Development, and Simulation of Everyday Mindreading* (pp. 253-266). Oxford: Basil Blackwell.

Premack, D., & Woodruff, G. (1978). Does the Chimpanzee Have a 'Theory of Mind'? *Behavioural and Brain Sciences; 4*, 515-526.

Pribram, K. H. (1990). From Metaphors to Models: The Use of Analogy in Neuropsychology. In D. E. Leary (Ed.), *Metaphors in the History of Psychology* (pp. 79-103). Cambridge: Cambridge University Press.

Putnam, H. (1975). The Meaning of 'Meaning'. In K. Gunderson (Ed.), *Language, Mind, and Knowledge.* Minneapolis: University of Minnesota Press.

Pylyshyn, Z. W. (1978). When is Attribution of Beliefs Justified? *Behavioral and Brain Sciences, 1*, 592-593.

Pylyshyn, Z. W. (1980). The 'Causal Power' of Machines. *Behavioral and Brain Sciences, 3*, 442-444.

Pylyshyn, Z. W. (1989). Computing in Cognitive Science. In M. I. Posner (Ed.), *Foundations of Cognitive Science* (pp. 49-92). Cambridge, Massachusetts: MIT Press.

Rao, A. S., & Georgeff, M. P. (1995). *BDI Agents: From Theory to Practice.* Paper presented at the the first International Conference on Multi-Agent Systems, ICMAS'95, San Fransisco, USA.

Ricard, M., & Allard, L. (1993) The Reaction of 9- and 10-Month-Old Infants to an Unfamiliar Animal. *Journal of Genetic Psychology, 154*(1), 5-16.

Ridley, M. (1993). *The Red Queen: Sex and the Evolution of Human Nature.* Viking.

Ristau, C. A. (1991). Before Mindreading: Attention, Purposes, and Deception in Birds. In A. Whiten (Ed.), *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading* (pp. 209-222). Oxford: Basil Blackwell.

Romanes, G. (1886). *Animal Intelligence.* (Fourth ed.). London: Kegan Paul, Trench, and Co.

Rorty, R. (1993). Consciousness, Intentionality, and Pragmatism. In S. M. Christiansen & D. R. Turner (Eds.), *Folk Psychology and the Philosophy of Mind* (p. 388-404). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Rosenschein, S. J. (1985). *Formal Theories of Knowledge in AI and Robotics* (Technical Report 362): SRI International, Menlo Park, CA.

Ryle, G. (1949). *The Concept of Mind.* Hutchinson.

Samet, J. (1993). Autism and Theory of Mind: Some Philosophical Perspectives. In S. Baron-Cohen, H. Tager-Flusberg, & D. J. Cohen (Eds.), *Understanding Other Minds: Perspectives from Autism* (pp. 427-449). Oxford: Oxford University Press.

Sandberg, J., & Wielinga, R. (1991). *How Situated is Cognition?* Paper presented at the twelfth International Joint Conference on Artificial Intelligence (IJCAI'91).

Schank, R. C., & Abelson, R. P. (1977). *Scripts, Plans, Goals, and Understanding.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Schank, R. C., & Seifert, C. M. (1985). Modeling Memory and Learning. In M. Shafto (Ed.), *How We Know* (pp. 60-88). New York: Harper and Row.

Searle, J., R. (1984). *Minds, Brains, and Science.* London: British Broadcasting Corporation.

Searle, J. R. (1969). *Speech Acts.* Cambridge: Cambridge University Press.

Searle, J. R. (1980). Minds, Brains, and Programs. *Behavioural and Brain Sciences, 3,* 417-424.

Searle, J. R. (1983). *Intentionality.* Cambridge: Cambridge University Press.

Searle, J. R. (1990). Consciousness, Explanatory Inversion, and Cognitive Science. *Behavioral and Brain Sciences, 13,* 585-642.

Searle, J. R. (1992). *The Rediscovery of the Mind.* Cambridge, Massachusetts: MIT Press.

Shoham, Y. (1992). Agent Oriented Programming. *Artificial Intelligence, 60*(1), 51-92.

Shultz, T. R. (1988). Assessing Intention: A Computational Model. In J. W. Astington, P. L. Harris, & D. R. Olson (Eds.), *Developing Theories of Mind* (pp. 341-367), Cambridge: Cambridge University Press.

Shultz, T. R. (1991). From Agency to Intention: A Rule-Based Computational Approach. In A. Whiten (Ed.), *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading* (pp. 79-95). Oxford: Basil Blackwell.

Simon, H. A. (1975). The Functional Equivalence of Problem Solving Skills. *Cognitive Science, 14,* 268-288.

Sloman, A. (1993). The Mind as a Control System. In C. Hookway & D. Peterson (Eds.), *Philosophy and Cognitive Science*. Cambridge: Cambridge University Press, Royal Institute of Philosophy Supplement 34.

Sloman, A. (1996). *What is it Like to be a Rock?* Unpublished web article available at URL: *http://www.cs.bham.ac.uk/~axs/misc/like_to_be_a_rock/:*

Sokal, R. R., & Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy*. San Francisco: W. H. Freeman.

Stalker, D. F. (1978). Why Machines Can't Think: A Reply to James Moor. *Philosophical Studies, 34*, 317-320.

Steels, L. (1988). *Steps Towards Common Sense*. Paper presented at the European Conference on Artificial Intelligence, ECAI'88, Munich.

Stich, S. (1983). *From Folk Psychology to Cognitive Science*. Cambridge, Massachusetts: MIT Press.

Stich, S., & Nichols, S. (1992). Folk Psychology: Simulation or Tacit Theory? *Mind and Language, 7*(1).

Suchman, L. A. (1987). *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge: Cambridge University Press.

Tamir, P., & Zohar, A. (1991). Anthropomorphism and Teleology in Reasoning about Biological Phenomena. *Science Education, 75*(1), 57-67.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind, LIX*(2236), 433-460.

Turkle, S. (1984). *The Second Self*. Simon and Schuster.

Turkle, S. (1988). Artifical Intelligence and Psychoanalysis: A New Alliance. *Daedalus, 117*(1), 241-268.

Tversky, A. (1977). Features of Similarity. *Psychological Review, 84*, 327-352.

Van Gulick, R. (1988). Consciousness, Intrinsic Intentionality, and Self-Understanding Machines. In A. J. Marcel & E. Bisiach (Eds.), *Consciousness in Contemporary Science* (pp. 78-100). Oxford: Oxford University Press.

Von Eckardt, B. (1993). *What is Cognitive Science?* Cambridge, Massachusetts: MIT Press.

Walker, J. H., Sproull, L., & Subramani, R. (1994). *Using a Human Face in an Interface*. Paper presented at CHI'94.

Wallis, C. (1992). Asymmetric Dependence and Mental Representation. *PSYCOLOQUY, 3*(70).

Weizenbaum, J. (1966, January 1966). ELIZA: A Computer Program for the Study of Natural Language Communication between man and machine. *Communications of the ACM, 9*, 36-45.

Weizenbaum, J. (1976). *Computer Power and Human Reason*. W. H. Freeman.

Wellman, H. M. (1990). *The Child's Theory of Mind*. Cambridge, Massachusetts: MIT Press.

Wellman, H. M. (1991). From Desires to Beliefs: Acquisition of a Theory of Mind. In A. Whiten (Ed.), *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading* (pp. 19-38). Oxford: Basil Blackwell.

Wesson, R., Hayes-Roth, F., Burge, J. W., Stasz, C., & Sunshine, C. A. (1981). Network Structures for Distributed Situation Assessment. *IEEE Transactions on Systems, Man, and Cybernetics, 11*(1), 5-23.

Whitby, B. (1996). The Turing Test: AI's Biggest Blind Alley? In P. Millican & A. Clark (Eds.), *Machines and Thought: The Legacy of Alan Turing* (Vol. 1, pp. 53-62). Oxford: Clarendon Press.

Whiten, A. (Ed.). (1991). *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading.* Oxford: Basil Blackwell.

Whiten, A., & Byrne, R. W. (1991). Human Ontogeny and Primate Phylogeny. In A. Whiten (Ed.), *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading* (pp. 267-282). Oxford: Basil Blackwell.

Wilkes, K. V. (1991). The Relationship Between Scientific Psychology and Common Sense Psychology. *Synthese, 89*, 15-39.

Wimmer, H., & Perner, J. (1983). Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception. *Cognition, 13*, 103-128.

Winograd, T. (1972). Understanding Natural Language. *Cognitive Psychology, 1*, 1-191.

Wittgenstein, L. (1953). *Philosophical Investigations.* Oxford: Basil Blackwell.

Woolgar, S. (1985). Why Not a Sociology of Machines? The Case of Sociology and Artificial Intelligence. *Sociology, 19*, 557-572.

Wundt, W. (1900-1920). *Völkerpsychologie.* (Vol. 1-10). London: Allen and Unwin.

Zaitchik, D. (1990). When Representations Conflict With Reality: The Preschooler's Problem With False Beliefs and 'False' Photographs. *Cognition, 35*(41-68).

Appendices

BLANK IN ORIGINAL

Appendix A

Models and traces for Baron-Cohen *et al.*'s false belief test

mfb.model

```
;;; Model definition for the reference version of Baron-Cohen et al.'s false belief test, described in
;;; chapters 9 and 10.

(load-file "basic.component")
(load-file "forms.component")

(in-model "FALSE-BELIEF" :clearp t)

(progn . #.(include-file "10.1.objects.freehand.text"))

(prove '(stop-trace-all))
(prove '(start-trace event))
(prove '(start-trace question))

(load-file "10.2.script.freehand.text")
(load-file "10.3.questions.freehand.text")
```

mfb.trace — *output from mfb.model*

```
; Loading file "mfb.model"...
;  Loading file "basic.component"...
;   Loading file "predicate.factor"...
;   Loading file "similarity.factor"...
;   Loading file "familiarity.factor"...
;   Loading file "animation.factor"...
;   Loading file "structure.factor"...
;   Including file "09.2.people.freehand.text"...
;   Including file "09.5.disintent.freehand.text"...
;   Including file "09.7.intent.freehand.text"...
;   Including file "09.4.disphys.freehand.text"...
;   Including file "09.6.phys.freehand.text"...
;  Loading file "forms.component"...
;  Including file "10.1.objects.freehand.text"...
;  Loading file "10.2.script.freehand.text"...
event basket (handle-event (perceived (put-in basket room)))
event basket (handle-event (perceived (place basket room)))
event box (handle-event (perceived (put-in box room)))
event basket (handle-event (perceived (put-in box room)))
event box (handle-event (perceived (place box room)))
event box (handle-event (perceived (place basket room)))
event marble (handle-event (perceived (put-in marble room)))
event box (handle-event (perceived (put-in marble room)))
event basket (handle-event (perceived (put-in marble room)))
event marble (handle-event (perceived (place marble room)))
event marble (handle-event (perceived (place box room)))
event marble (handle-event (perceived (place basket room)))
event sally (handle-event (perceived (put-in sally room)))
```

```
event marble (handle-event (perceived (put-in sally room)))
event box (handle-event (perceived (put-in sally room)))
event basket (handle-event (perceived (put-in sally room)))
event sally (handle-event (perceived (place sally room)))
event sally (handle-event (perceived (place marble room)))
event sally (handle-event (perceived (place box room)))
event sally (handle-event (perceived (place basket room)))
event anne (handle-event (perceived (put-in anne room)))
event sally (handle-event (perceived (put-in anne room)))
event marble (handle-event (perceived (put-in anne room)))
event box (handle-event (perceived (put-in anne room)))
event basket (handle-event (perceived (put-in anne room)))
event anne (handle-event (perceived (place anne room)))
event anne (handle-event (perceived (place sally room)))
event anne (handle-event (perceived (place marble room)))
event anne (handle-event (perceived (place box room)))
event anne (handle-event (perceived (place basket room)))
event alison (handle-event (perceived (put-in alison room)))
event anne (handle-event (perceived (put-in alison room)))
event sally (handle-event (perceived (put-in alison room)))
event marble (handle-event (perceived (put-in alison room)))
event box (handle-event (perceived (put-in alison room)))
event basket (handle-event (perceived (put-in alison room)))
event alison (handle-event (perceived (place alison room)))
event alison (handle-event (perceived (place anne room)))
event alison (handle-event (perceived (place sally room)))
event alison (handle-event (perceived (place marble room)))
event alison (handle-event (perceived (place box room)))
event alison (handle-event (perceived (place basket room)))
event marble (handle-event (perceived (put-in marble basket)))
event alison (handle-event (perceived (put-in marble basket)))
event anne (handle-event (perceived (put-in marble basket)))
event sally (handle-event (perceived (put-in marble basket)))
event box (handle-event (perceived (put-in marble basket)))
event basket (handle-event (perceived (put-in marble basket)))
event marble (handle-event (perceived (place marble basket)))
event alison (handle-event (perceived (take-out sally room)))
event anne (handle-event (perceived (take-out sally room)))
event box (handle-event (perceived (take-out sally room)))
event basket (handle-event (perceived (take-out sally room)))
event marble (handle-event (perceived (put-in marble box)))
event alison (handle-event (perceived (put-in marble box)))
event anne (handle-event (perceived (put-in marble box)))
event box (handle-event (perceived (put-in marble box)))
event basket (handle-event (perceived (put-in marble box)))
event marble (handle-event (perceived (place marble box)))
event sally (handle-event (perceived (put-in sally room)))
event alison (handle-event (perceived (put-in sally room)))
event anne (handle-event (perceived (put-in sally room)))
event box (handle-event (perceived (put-in sally room)))
event basket (handle-event (perceived (put-in sally room)))
event sally (handle-event (perceived (place sally room)))
event sally (handle-event (perceived (place alison room)))
event sally (handle-event (perceived (place anne room)))
event sally (handle-event (perceived (place box room)))
event sally (handle-event (perceived (place basket room)))
;  Loading file "10.3.questions.freehand.text"...
question alison (believes alison (inside marble ?where))
yes alison believes (inside marble box)
question alison (believes sally (inside marble ?where))
yes sally believes (inside marble basket)
question alison (believes anne (inside marble ?where))
yes anne believes (inside marble box)
```

mfbsimulationfail.model

;;; Model definition for the first version of the simulation theory for Baron-Cohen *et al*'s false belief
;;; test, described in chapter 10. This version fails the false belief test.

```
(load-file "basic.component")
(load-file "forms.component")

(in-model "FALSE-BELIEF" :clearp t)

(progn . #.(include-file "10.1.objects.freehand.text"))

(model-stance simulation-fail-intentional-stance
  :built-on 'basic-intentional-stance
  :clauses '#.(include-file "10.4.simfail.freehand.text"))

(model-object alison :form person
  :stances (simulation-fail-intentional-stance
            basic-physical-stance)
  :database basic-person-object-database)

(prove '(stop-trace-all))
(prove '(start-trace question))

(load-file "10.2.script.freehand.text")
(load-file "10.3.questions.freehand.text")
```

mfbsimulationfail.trace — *output from mfbsimulationfail.model*

```
; Loading file "mfbsimulationfail.model"...
;  Loading file "basic.component"...
;   Loading file "predicate.factor"...
;   Loading file "similarity.factor"...
;   Loading file "familiarity.factor"...
;   Loading file "animation.factor"...
;   Loading file "structure.factor"...
;   Including file "09.2.people.freehand.text"...
;   Including file "09.5.disintent.freehand.text"...
;   Including file "09.7.intent.freehand.text"...
;   Including file "09.4.disphys.freehand.text"...
;   Including file "09.6.phys.freehand.text"...
;  Loading file "forms.component"...
;  Including file "10.1.objects.freehand.text"...
;  Including file "10.4.simfail.freehand.text"...
;  Loading file "10.2.script.freehand.text"...
;  Loading file "10.3.questions.freehand.text"...
question alison (believes alison (inside marble ?where))
yes alison believes (inside marble box)
question alison (believes sally (inside marble ?where))
no sally does not believe (inside marble ?where)
question alison (believes anne (inside marble ?where))
no anne does not believe (inside marble ?where)
```

mfbsimulation.model

```
;;; Model definition for the second version of the simulation theory for Baron-Cohen et al.'s false
;;; belief test, described in chapter 10.  This version passes the false belief test.

(load-file "basic.component")
(load-file "forms.component")

(in-model "FALSE-BELIEF" :clearp t)

(progn . #.(include-file "10.1.objects.freehand.text"))

(model-stance simulation-intentional-stance
  :built-on 'basic-intentional-stance
  :clauses '#.(append (include-file "10.5.sim.freehand.text")
                      (include-file "10.4.simfail.freehand.text")))

(model-object alison :form person
  :stances (simulation-intentional-stance
            basic-physical-stance)
  :database basic-person-object-database)

(prove '(stop-trace-all))
(prove '(start-trace question))

(load-file "10.2.script.freehand.text")
(load-file "10.3.questions.freehand.text")
```

mfbsimulation.trace — *output from mfbsimulation.model*

```
; Loading file "mfbsimulation.model"...
;  Loading file "basic.component"...
;   Loading file "predicate.factor"...
;   Loading file "similarity.factor"...
;   Loading file "familiarity.factor"...
;   Loading file "animation.factor"...
;   Loading file "structure.factor"...
;   Including file "09.2.people.freehand.text"...
;   Including file "09.5.disintent.freehand.text"...
;   Including file "09.7.intent.freehand.text"...
;   Including file "09.4.disphys.freehand.text"...
;   Including file "09.6.phys.freehand.text"...
;  Loading file "forms.component"...
;  Including file "10.1.objects.freehand.text"...
;  Including file "10.5.sim.freehand.text"...
;  Including file "10.4.simfail.freehand.text"...
;  Loading file "10.2.script.freehand.text"...
;  Loading file "10.3.questions.freehand.text"...
question alison (believes alison (inside marble ?where))
yes alison believes (inside marble box)
question alison (believes sally (inside marble ?where))
yes sally believes (inside marble basket)
question alison (believes anne (inside marble ?where))
yes anne believes (inside marble box)
```

mfbcopy.model

;;; Model definition for Chandler's copy theory for Baron-Cohen *et al.*'s false belief test, described in
;;; chapter 10. This version fails the false belief test.

```
(load-file "basic.component")
(load-file "forms.component")

(in-model "FALSE-BELIEF" :clearp t)

(progn . #.(include-file "10.1.objects.freehand.text"))

(model-stance copy-intentional-stance
  :built-on 'basic-intentional-stance
  :clauses '#.(include-file "10.6.copy1.freehand.text"))

(model-object-database copy-person-object-database
  :built-on 'basic-person-object-database
  :clauses '#.(include-file "10.7.copy2.freehand.text"))

(model-object alison :form person
  :stances (copy-intentional-stance
            basic-physical-stance)
  :database copy-person-object-database)

(prove '(stop-trace-all))
(prove '(start-trace question))

(load-file "10.2.script.freehand.text")
(load-file "10.3.questions.freehand.text")
```

mfbcopy.trace — *output from mfbcopy.model*

```
; Loading file "mfbcopy.model"...
;  Loading file "basic.component"...
;   Loading file "predicate.factor"...
;   Loading file "similarity.factor"...
;   Loading file "familiarity.factor"...
;   Loading file "animation.factor"...
;   Loading file "structure.factor"...
;   Including file "09.2.people.freehand.text"...
;   Including file "09.5.disintent.freehand.text"...
;   Including file "09.7.intent.freehand.text"...
;   Including file "09.4.disphys.freehand.text"...
;   Including file "09.6.phys.freehand.text"...
;  Loading file "forms.component"...
;  Including file "10.1.objects.freehand.text"...
;  Including file "10.6.copy1.freehand.text"...
;  Including file "10.7.copy2.freehand.text"...
;  Loading file "10.2.script.freehand.text"...
;  Loading file "10.3.questions.freehand.text"...
question alison (believes alison (inside marble ?where))
yes alison believes (inside marble box)
question alison (believes sally (inside marble ?where))
no sally does not believe (inside marble ?where)
question alison (believes anne (inside marble ?where))
no anne does not believe (inside marble ?where)
```

mfbsituationfail.model

;;; Model definition for the first version of Perner's situation theory for Baron-Cohen *et al.*'s false
;;; belief test, described in chapter 10. This version fails the false belief test.

```
(load-file "basic.component")
(load-file "forms.component")

(in-model "FALSE-BELIEF" :clearp t)

(progn . #.(include-file "10.1.objects.freehand.text"))

(model-stance situation-intentional-stance
  :class 'intentional
  :built-on 'basic-intentional-stance
  :clauses '#.(append (include-file "10.8.sitascr.freehand.text")
                      (include-file "10.9.sitfail.freehand.text")))

(model-object alison :form person
  :stances (situation-intentional-stance
            basic-physical-stance)
  :database basic-person-object-database)

(prove '(stop-trace-all))
(prove '(start-trace question))

(load-file "10.2.script.freehand.text")
(load-file "10.3.questions.freehand.text")
```

mfbsituationfail.trace — *output from mfbsituationfail.model*

```
; Loading file "mfbsituationfail.model"...
;  Loading file "basic.component"...
;   Loading file "predicate.factor"...
;   Loading file "similarity.factor"...
;   Loading file "familiarity.factor"...
;   Loading file "animation.factor"...
;   Loading file "structure.factor"...
;   Including file "09.2.people.freehand.text"...
;   Including file "09.5.disintent.freehand.text"...
;   Including file "09.7.intent.freehand.text"...
;   Including file "09.4.disphys.freehand.text"...
;   Including file "09.6.phys.freehand.text"...
;  Loading file "forms.component"...
;  Including file "10.1.objects.freehand.text"...
;  Including file "10.8.sitascr.freehand.text"...
;  Including file "10.9.sitfail.freehand.text"...
;  Loading file "10.2.script.freehand.text"...
;  Loading file "10.3.questions.freehand.text"...
question alison (believes alison (inside marble ?where))
yes alison believes (inside marble box)
question alison (believes sally (inside marble ?where))
yes sally believes (inside marble box)
question alison (believes anne (inside marble ?where))
yes anne believes (inside marble box)
```

mfbsituation.model

```
;;; Model definition for the second version of Perner's situation theory for Baron-Cohen et al.'s false
;;; belief test, described in chapter 10.  This version passes the false belief test.

(load-file "basic.component")
(load-file "forms.component")

(in-model "FALSE-BELIEF" :clearp t)

(progn . #.(include-file "10.1.objects.freehand.text"))

(model-stance situation-intentional-stance
  :class 'intentional
  :built-on 'basic-intentional-stance
  :clauses '#.(append (include-file "10.8.sitascr.freehand.text")
                      (include-file "10.10.sit.freehand.text")
                      (include-file "10.9.sitfail.freehand.text")))

(model-object alison :form person
  :stances (situation-intentional-stance
            basic-physical-stance)
  :database basic-person-object-database)

(prove '(stop-trace-all))
(prove '(start-trace question))

(load-file "10.2.script.freehand.text")
(load-file "10.3.questions.freehand.text")
```

mfbsituation.trace — *output from mfbsituation.model*

```
; Loading file "mfbsituation.model"...
;  Loading file "basic.component"...
;   Loading file "predicate.factor"...
;   Loading file "similarity.factor"...
;   Loading file "familiarity.factor"...
;   Loading file "animation.factor"...
;   Loading file "structure.factor"...
;   Including file "09.2.people.freehand.text"...
;   Including file "09.5.disintent.freehand.text"...
;   Including file "09.7.intent.freehand.text"...
;   Including file "09.4.disphys.freehand.text"...
;   Including file "09.6.phys.freehand.text"...
;  Loading file "forms.component"...
;  Including file "10.1.objects.freehand.text"...
;  Including file "10.8.sitascr.freehand.text"...
;  Including file "10.10.sit.freehand.text"...
;  Including file "10.9.sitfail.freehand.text"...
;  Loading file "10.2.script.freehand.text"...
;  Loading file "10.3.questions.freehand.text"...
question alison (believes alison (inside marble ?where))
yes alison believes (inside marble box)
question alison (believes sally (inside marble ?where))
yes sally believes (inside marble basket)
question alison (believes anne (inside marble ?where))
yes anne believes (inside marble box)
```

Appendix B

Models and traces for Searle's Chinese Room

mcr.model

;;; Model definition for the basic version of Searle's Chinese Room, described in chapter 12. Note
;;; that all these models also use specialised script and database components for the Chinese Room,
;;; described in Appendix C, files *mcrscripts.component* and *mcrdatabases.component*.

```
(load-file "basic.component")
(load-file "forms.component")

(in-model "CHINESE-ROOM" :clearp t)

(load-file "mcrscripts.component")
(load-file "mcrdatabases.component")

(load-file "12.1.objects.freehand.text")

(prove '(stop-trace-all))

(load-file "12.2.script.freehand.text")
```

mcr.trace — *output from mcr.model*

```
; Loading file "mcr.model"...
;  Loading file "basic.component"...
;   Loading file "predicate.factor"...
;   Loading file "similarity.factor"...
;   Loading file "familiarity.factor"...
;   Loading file "animation.factor"...
;   Loading file "structure.factor"...
;   Including file "09.2.people.freehand.text"...
;   Including file "09.5.disintent.freehand.text"...
;   Including file "09.7.intent.freehand.text"...
;   Including file "09.4.disphys.freehand.text"...
;   Including file "09.6.phys.freehand.text"...
;  Loading file "forms.component"...
;  Loading file "mcrscripts.component"...
;  Loading file "mcrdatabases.component"...
;  Loading file "12.1.objects.freehand.text"...
;  Loading file "12.2.script.freehand.text"...
searle cant take any stance to chinese-room for
  (perceived chinese-room (put-in searle room))
searle cant take any stance to chinese-room for
  (perceived chinese-room (place searle room))
searle cant take any stance to chinese-room for
  (perceived chinese-room (place chinese-room room))
searle cant take any stance to chinese-room for
  (perceived chinese-room (place alison room))
yes alison believes (inside searle room)
searle cant take any stance to chinese-room for
  (believes chinese-room (inside ?something ?where))
```

```
writing script number 1
(model-object gadget :form marble)
(model-object widget :form marble)
(model-object doodah :form cupboard)
(model-object thingamy :form box)
(model-object sally :form doll)
(model-object alison :form person
  :stances (basic-intentional-stance basic-physical-stance)
  :database basic-person-object-database)
(tell-model (put-in gadget room))
(tell-model (put-in widget room))
(tell-model (put-in doodah room))
(tell-model (put-in thingamy room))
(tell-model (put-in sally room))
(tell-model (put-in alison room))
(tell-model (put-in gadget doodah))
(tell-model (take-out sally room))
(tell-model (take-out gadget doodah))
(tell-model (put-in gadget thingamy))
(tell-model (put-in sally room))

telling the script to alison

now asking questions
ask-in-model alison (believes sally (inside gadget doodah))
yes sally believes (inside gadget doodah)
my answer
yes sally believes (inside gadget doodah)
the answer is correct
ask-in-model alison (believes alison (inside gadget doodah))
no alison does not believe (inside gadget doodah)
my answer
no alison does not believe (inside gadget doodah)
the answer is correct
rating for alison of form person is 1.0

telling the script to chinese-room

now asking questions
ask-in-model chinese-room (believes sally (inside gadget doodah))
yes sally believes (inside gadget doodah)
my answer
yes sally believes (inside gadget doodah)
the answer is correct
ask-in-model chinese-room (believes alison (inside gadget doodah))
no alison does not believe (inside gadget doodah)
my answer
no alison does not believe (inside gadget doodah)
the answer is correct
rating for chinese-room of form room is 1.0
```

;;; 9 similar script tests omitted here.

```
yes alison believes (inside searle room)
no chinese-room does not believe (inside ?something ?where)
searle cant take any stance to chinese-room for
  (perceived chinese-room (put-in searle-in-the-room chinese-room))
yes alison believes (inside searle room)
searle cant take any stance to chinese-room for
  (believes chinese-room (inside ?something ?where))
no searle-in-the-room does not believe (inside ?something ?where)
```

mcrsystems.model

```
;;; Model definition for the systems reply version of Searle's Chinese Room, described in chapter 12.

(load-file "basic.component")
(load-file "forms.component")

(in-model "CHINESE-ROOM" :clearp t)

(load-file "mcrscripts.component")
(load-file "mcrdatabases.component")
(load-file "12.1.objects.freehand.text")

(prove '(stop-trace-all))

(load-file "12.3.systems.freehand.text")

(load-file "12.2.script.freehand.text")
```

mcrsystems.trace — *output from mcrsystems.model*

```
; Loading file "mcrsystems.model"...
;  Loading file "basic.component"...
;   Loading file "predicate.factor"...
;   Loading file "similarity.factor"...
;   Loading file "familiarity.factor"...
;   Loading file "animation.factor"...
;   Loading file "structure.factor"...
;   Including file "09.2.people.freehand.text"...
;   Including file "09.5.disintent.freehand.text"...
;   Including file "09.7.intent.freehand.text"...
;   Including file "09.4.disphys.freehand.text"...
;   Including file "09.6.phys.freehand.text"...
;  Loading file "forms.component"...
;  Loading file "mcrscripts.component"...
;  Loading file "mcrdatabases.component"...
;  Loading file "12.1.objects.freehand.text"...
;  Loading file "12.3.systems.freehand.text"...
;  Loading file "12.2.script.freehand.text"...
yes alison believes (inside searle room)
yes chinese-room believes (inside searle room)

;;; 10 script tests omitted here.

yes alison believes (inside searle room)
yes chinese-room believes (inside searle room)
yes alison believes (inside searle room)
yes chinese-room believes (inside searle room)
no searle-in-the-room does not believe (inside ?something ?where)
```

mcrrobot.model

;;; Model definition for the robot reply version of Searle's Chinese Room, described in chapter 12.

```
(load-file "basic.component")
(load-file "forms.component")

(in-model "CHINESE-ROOM" :clearp t)

(load-file "mcrscripts.component")
(load-file "mcrdatabases.component")

(load-file "12.1.objects.freehand.text")

(prove '(stop-trace-all))

(load-file "12.4.robot.freehand.text")

(load-file "12.2.script.freehand.text")
```

mcrrobot.trace — *output from mcrrobot.model*

```
; Loading file "mcrrobot.model"...
;  Loading file "basic.component"...
;   Loading file "predicate.factor"...
;   Loading file "similarity.factor"...
;   Loading file "familiarity.factor"...
;   Loading file "animation.factor"...
;   Loading file "structure.factor"...
;   Including file "09.2.people.freehand.text"...
;   Including file "09.5.disintent.freehand.text"...
;   Including file "09.7.intent.freehand.text"...
;   Including file "09.4.disphys.freehand.text"...
;   Including file "09.6.phys.freehand.text"...
;  Loading file "forms.component"...
;  Loading file "mcrscript.component"...
;  Loading file "mcrdatabases.component"...
;  Loading file "12.1.objects.freehand.text"...
;  Loading file "12.4.robot.freehand.text"...
;  Loading file "12.2.script.freehand.text"...
yes alison believes (inside searle room)
yes chinese-room believes (inside searle room)
```

;;; 10 script tests omitted here.

```
yes alison believes (inside searle room)
yes chinese-room believes (inside searle room)
yes alison believes (inside searle room)
yes chinese-room believes (inside searle room)
no searle-in-the-room does not believe (inside ?something ?where)
```

mcrneural.model

```
;;; Model definition for the neural reply version of Searle's Chinese Room, described in chapter 12.

(load-file "basic.component")
(load-file "forms.component")

(in-model "CHINESE-ROOM" :clearp t)

(load-file "mcrscripts.component")
(load-file "mcrdatabases.component")

(defconstant neuron-count 10)
(defun get-neuron-name (number)
  (intern (format () "ROOM-NEURON-~D" number)))

(load-file "12.1.objects.freehand.text")

(prove '(stop-trace-all))

(loop for number from 1 to neuron-count
      for name = (get-neuron-name number)
      do (eval `(model-object ,name :form neuron)))

(tell-model (put-in alison room))
(tell-model (put-in chinese-room room))

(loop for number from 1 to neuron-count
      for name = (get-neuron-name number)
      do (eval `(progn (tell-model (put-in ,name room))
                       (tell-model (put-in ,name chinese-room)))))

(tell-model (put-in searle room))

(ask-object-if searle (believes alison (inside ?something ?where)))
(ask-object-if searle (believes chinese-room (inside ?something ?where)))

(ask-object-to searle (do-turing-test))

(ask-object-if searle (believes alison (inside ?something ?where)))
(ask-object-if searle (believes chinese-room (inside ?something ?where)))

(loop for number from 1 to neuron-count
      for name = (get-neuron-name number)
      do (eval `(progn (tell-model (take-out ,name chinese-room))
                       (tell-model (put-in ,name chinese-room)))))

(ask-object-if searle (believes alison (inside ?something ?where)))
(ask-object-if searle (believes chinese-room (inside ?something ?where)))
(ask-object-if searle (believes searle-in-the-room (inside ?something ?where)))
```

mcrneural.trace — *output from mcrneural.model*

```
; Loading file "mcrneural.model"...
;  Loading file "basic.component"...
;   Loading file "predicate.factor"...
;   Loading file "similarity.factor"...
;   Loading file "familiarity.factor"...
;   Loading file "animation.factor"...
;   Loading file "structure.factor"...
;   Including file "09.2.people.freehand.text"...
;   Including file "09.5.disintent.freehand.text"...
;   Including file "09.7.intent.freehand.text"...
;   Including file "09.4.disphys.freehand.text"...
;   Including file "09.6.phys.freehand.text"...
;  Loading file "forms.component"...
```

```
;  Loading file "mcrscripts.component"...
;  Loading file "mcrdatabases.component"...
;  Loading file "12.1.objects.freehand.text"...
searle cant take any stance to chinese-room for
  (perceived chinese-room (put-in searle room))
searle cant take any stance to chinese-room for
  (perceived chinese-room (place searle room))
searle cant take any stance to chinese-room for
  (perceived chinese-room (place chinese-room room))
searle cant take any stance to chinese-room for
  (perceived chinese-room (place alison room))
yes alison believes (inside searle room)
searle cant take any stance to chinese-room for
  (believes chinese-room (inside ?something ?where))
```

;;; 10 script tests omitted here.

```
yes alison believes (inside searle room)
no chinese-room does not believe (inside ?something ?where)
searle cant take any stance to room-neuron-1 for
  (perceived room-neuron-1 (take-out room-neuron-1 chinese-room))
searle cant take any stance to chinese-room for
  (perceived chinese-room (put-in room-neuron-1 chinese-room))
searle cant take any stance to room-neuron-2 for
  (perceived room-neuron-2 (take-out room-neuron-2 chinese-room))
searle cant take any stance to chinese-room for
  (perceived chinese-room (take-out room-neuron-2 chinese-room))
searle cant take any stance to chinese-room for
  (perceived chinese-room (put-in room-neuron-2 chinese-room))
searle cant take any stance to room-neuron-3 for
  (perceived room-neuron-3 (take-out room-neuron-3 chinese-room))
searle cant take any stance to chinese-room for
  (perceived chinese-room (take-out room-neuron-3 chinese-room))
searle cant take any stance to chinese-room for
  (perceived chinese-room (put-in room-neuron-3 chinese-room))
searle cant take any stance to room-neuron-4 for
  (perceived room-neuron-4 (take-out room-neuron-4 chinese-room))
searle cant take any stance to chinese-room for
  (perceived chinese-room (take-out room-neuron-4 chinese-room))
searle cant take any stance to chinese-room for
  (perceived chinese-room (put-in room-neuron-4 chinese-room))
searle cant take any stance to room-neuron-5 for
  (perceived room-neuron-5 (take-out room-neuron-5 chinese-room))
searle cant take any stance to chinese-room for
  (perceived chinese-room (take-out room-neuron-5 chinese-room))
searle cant take any stance to chinese-room for
  (perceived chinese-room (put-in room-neuron-5 chinese-room))
searle cant take any stance to room-neuron-6 for
  (perceived room-neuron-6 (take-out room-neuron-6 chinese-room))
searle cant take any stance to chinese-room for
  (perceived chinese-room (take-out room-neuron-6 chinese-room))
searle cant take any stance to chinese-room for
  (perceived chinese-room (put-in room-neuron-6 chinese-room))
searle cant take any stance to room-neuron-7 for
  (perceived room-neuron-7 (take-out room-neuron-7 chinese-room))
searle cant take any stance to chinese-room for
  (perceived chinese-room (take-out room-neuron-7 chinese-room))
searle cant take any stance to room-neuron-8 for
  (perceived room-neuron-8 (take-out room-neuron-8 chinese-room))
searle cant take any stance to room-neuron-9 for
  (perceived room-neuron-9 (take-out room-neuron-9 chinese-room))
searle cant take any stance to room-neuron-10 for
  (perceived room-neuron-10 (take-out room-neuron-10 chinese-room))
yes alison believes (inside searle room)
yes chinese-room believes (inside room-neuron-7 chinese-room)
no searle-in-the-room does not believe (inside ?something ?where)
```

Appendix C
Model component listings

basic.component

```
;;; The basic stance component, shared by all the different models.  These basic stances can be
;;; overridden by the different models as and when required.  This allows us to share most of the
;;; implementation of the different models without having to replicate everything everywhere.

(load-file "predicate.factor")
(load-file "similarity.factor")
(load-file "familiarity.factor")
(load-file "animation.factor")
(load-file "structure.factor")

;;; Dividing classes and instances here is not all that simple.  We'd like to be able to use instances
;;; directly, but also to get the benefits of inheritance.  Thus, stances use a  simple prototype model,
;;; and the object database system.  That way, one stance can 'override' all or part of another simply
;;; by adding a few extra definitions.
;;;
;;; This object database is the default for all objects.  All objects will use this object database unless
;;; explicitly overridden in the object definition.

(model-object-database basic-object-database
  :clauses
    '(((results no ?stance-to ?action ?situation ^situation) :-
       (self ?self)
       (write-list (?self cant take any stance to ?stance-to for))
       (nl) (spaces 2) (write ?action)
       (cut))))

;;; This object database is the default for all people.  It includes the rules for people defined in
;;; figure 9.3.

(model-object-database basic-person-object-database
  :built-on 'basic-object-database
  :clauses '#.(include-file "09.3.people.freehand.text"))

(defpredicate get-situation
  ((get-situation ^situation) :-
   (situation ?situation))
  ((get-situation ())))

(defpredicate put-situation
  ((put-situation ?situation) :-
   (situation ?old-situation)
   (retract ((situation ?old-situation)))
   (asserta ((situation ?situation))))
  ((put-situation ?situation) :-
   (asserta ((situation ?situation)))))

;;; We use transform functions to encode the factors that we use for the system.  This makes it rather
```

```
;;; easier to look at and compare the factors used in balancing the different factors.

(defun transform (x &key scale translation)
  (+ (* scale x) translation))

;;; Define the basic physical and intentional stances, using the dispositions in figures 9.5 and 9.6, and
;;; the rules in figures 9.7 and 9.8.

(model-stance basic-intentional-stance
  :class 'intentional
  :dispositions '(predicate similarity familiarity animation structure context)
  :disposition-energy #.(first (include-file "09.6.disintent.freehand.text"))
  :clauses '#.(include-file "09.8.intent.freehand.text"))

(model-stance basic-physical-stance
  :class 'physical
  :dispositions '(predicate context)
  :disposition-energy #.(first (include-file "09.5.disphys.freehand.text"))
  :clauses '#.(include-file "09.7.phys.freehand.text"))
```

forms.component

```
;;; This rather dull model component simply contains the model's form definitions in the proper
;;; syntax.  These values and properties are all defined in figure 9.3.

(clear-forms)

(define-form-property has-head?
  (:type t)
  (:category similarity))

(define-form-property has-limbs?
  (:type t)
  (:category similarity))

(define-form-property solid?
  (:type t)
  (:category similarity))

(define-form-property width
  (:type real)
  (:category similarity))

(define-form-property height
  (:type real)
  (:category similarity))

(define-form-property depth
  (:type real)
  (:category similarity))

(define-form-property animated?
  (:type t)
  (:category animation))

(define-form person
  ((has-head? :value t)
   (has-limbs? :value t)
   (solid? :value t)
   (width :value 0.25)
   (height :value 1.0)
   (depth :value 0.125)
   (animated? :value t)))
```

```
(define-form doll
  ((has-head? :value t)
   (has-limbs? :value t)
   (solid? :value t)
   (width :value 0.05)
   (height :value 0.2)
   (depth :value 0.025)
   (animated? :value nil)))

(define-form robot
  ((has-head? :value t)
   (has-limbs? :value t)
   (solid? :value t)
   (width :value 0.25)
   (height :value 1.0)
   (depth :value 0.125)
   (animated? :value nil)))

(define-form room
  ((has-head? :value nil)
   (has-limbs? :value nil)
   (solid? :value nil)
   (width :value 4.0)
   (height :value 3.0)
   (depth :value 4.0)
   (animated? :value nil)))

(define-form cupboard
  ((has-head? :value nil)
   (has-limbs? :value nil)
   (solid? :value nil)
   (width :value 1.0)
   (height :value 2.0)
   (depth :value 0.5)
   (animated? :value nil)))

(define-form marble
  ((has-head? :value nil)
   (has-limbs? :value nil)
   (solid? :value t)
   (width :value 0.01)
   (height :value 0.01)
   (depth :value 0.01)
   (animated? :value nil)))

(define-form neuron
  ((has-head? :value nil)
   (has-limbs? :value nil)
   (solid? :value t)
   (width :value 0.00001)
   (height :value 0.00001)
   (depth :value 0.00001)
   (animated? :value nil)))

(define-form box
  ((has-head? :value nil)
   (has-limbs? :value nil)
   (solid? :value nil)
   (width :value 0.03)
   (height :value 0.02)
   (depth :value 0.02)
   (animated? :value nil)))

(define-form basket
  ((has-head? :value nil)
   (has-limbs? :value nil)
   (solid? :value nil)
   (width :value 0.03)
   (height :value 0.02)
   (depth :value 0.02)
   (animated? :value nil)))
```

```
(define-form chocolate
  ((has-head? :value nil)
   (has-limbs? :value nil)
   (solid? :value nil)
   (width :value 0.07)
   (height :value 0.02)
   (depth :value 0.005)
   (animated? :value nil)))
```

mcrscripts.component

;;; Script generator for the Chinese Room. We here define a generator for random experiments of
;;; the kind which will involve the false belief ascriptions. All these scripts involve a cast of objects
;;; and people, and all are set inside a room. Some objects are containers. Objects can be moved in
;;; and out of objects and people can be moved in and out of the room. Every story involves at least
;;; one person, at least one object, and at least one container.

```
(defpredicate forms
  ((forms doll (doll)))
  ((forms object (marble chocolate)))
  ((forms container (basket box cupboard))))

(defpredicate names
  ((names doll (sally anne maxi john mary)))
  ((names object (doobry gadget widget)))
  ((names container (thingamy gizmo doodah))))

(defpredicate get-object
  ((get-object ?class (model-object ^name :form ^form)) :-
   (forms ?class ?forms)
   (names ?class ?names)
   (anyof ?form ?forms)
   (anyof ?name ?names)))

(defpredicate get-cast
  ((get-cast (^object-objects ^container-objects ^doll-objects)) :-
   (random ?object-count 1 3)
   (random ?container-count 2 3)
   (random ?doll-count 1 3)
   (get-objects object ?object-count ?object-objects)
   (get-objects container ?container-count ?container-objects)
   (get-objects doll ?doll-count ?doll-objects)))

(defpredicate already-defined
  ((already-defined (model-object ?name :form ?form1)
     ((model-object ?name :form ?form2) . ?z)))
  ((already-defined ?x (?y . ?z)) :-
   (already-defined ?x ?z)))

(defpredicate get-objects
  ((get-objects ?class 0 ()))
  ((get-objects ?class ?count (^object . ^rest)) :-
   (is ?newcount (- '?count 1))
   (get-objects ?class ?newcount ?rest)
   (get-object ?class ?object)
   (not (already-defined ?object ?rest)))
  ((get-objects ?class ?count ^objects) :-
   (get-objects ?class ?count ?objects)))
```

;;; Having derived the cast, we can now define the script of movements. This will involve, at the
;;; least, moving one of the dolls into and out of the room. Other movements can only apply to
;;; containers.

```
(defpredicate object-form
  ((object-form ?kind ^form) :-
   (forms ?kind ?forms)
   (member ?form ?forms)))
```

;;; We now need to move things around a bit, before we can proceed to ask questions. This works
;;; by picking a few objects, which might be dolls, and moving them in containers and out of the
;;; room, and then finally merging these moves. We should also ensure that some things stay where
;;; they are because otherwise we'll end up right back where we started. This is actually rather
;;; complicated to define as a generator, because even when we adhere to the physical rules we need
;;; some constraints which require complicated temporal reasoning. This is needed to ensure that
;;; people could have acquired false beliefs. For now, just use a selection of segments of scripts.

```
(defpredicate get-movement-script
  ((get-movement-script ^script
     (((model-object ?object :form ?object-form) . ?rest-objects)
      ((model-object ?container :form ?container-form)
       (model-object ?other-container :form ?other-container-form) . ?rest-containers)
      ((model-object ?doll :form doll) . ?rest-dolls))) :-
   (anyof ?script ((((put-in ?object ?container)
                     (take-out ?doll room)
                     (take-out ?object ?container)
                     (put-in ?object ?other-container)
                     (put-in ?doll room))
                    ((believes ?doll (inside ?object ?container))
                     (believes alison (inside ?object ?container))))
                   (((put-in ?object ?container)
                     (take-out ?doll room)
                     (put-in ?doll room))
                    ((believes ?doll (inside ?object ?container))
                     (believes alison (inside ?object ?container))))
                   (((take-out ?doll room)
                     (put-in ?object ?container)
                     (put-in ?doll room))
                    ((believes ?doll (inside ?object room))
                     (believes alison (inside ?object room))))
                   (((put-in ?object ?container)
                     (take-out ?doll room)
                     (take-out ?object ?container)
                     (put-in ?object ?other-container))
                    ((believes ?doll (inside ?object ?container))
                     (believes alison (inside ?object ?container))))
                   (((put-in ?object ?container)
                     (take-out ?doll room))
                    ((believes ?doll (inside ?object ?container))
                     (believes alison (inside ?object ?container))))
                   (((take-out ?doll room)
                     (put-in ?object ?container))
                    ((believes ?doll (inside ?object room))
                     (believes alison (inside ?object room)))))))))

(defpredicate get-room-movements
  ((get-room-movements () ()))
  ((get-room-movements ((model-object ?name . ?options) . ?rest1)
     ((tell-model (put-in ^name room)) . ^rest2)) :-
   (get-room-movements ?rest1 ?rest2)))

(defpredicate get-script
  ((get-script (^script ^script-questions)) :-
   (get-cast (?objects ?containers ?dolls))
   (get-movement-script (?movements ?script-questions) (?objects ?containers ?dolls))
   (requote ?movement ?movements (tell-model ?movement) ?script-movements)
   (append ?objects ?containers ?list1)
   (append ?list1 ?dolls ?list2)
   (append ?list2 ((model-object alison :form person
                    :stances (basic-intentional-stance
                              basic-physical-stance)
                    :database basic-person-object-database)) ?all-objects)
```

```
(get-room-movements ?all-objects ?room-movements)
(append ?all-objects ?room-movements ?list3)
(append ?list3 ?script-movements ?script)))
```

mcrdatabases.component

```
;;; This component provides modified object databases which give Searle, Alison, and the Chinese
;;; Room the mental models needed so they can deal with the false belief test scripts.  Searle is also
;;; given rules which allow him to carry out and evaluate the Turing test needed to study the effects
;;; of familiarity.

(model-object-database modelled-person-object-database
  :built-on 'basic-person-object-database
  :clauses
    '(((clear) :-
       (model-name ?model)
       (in-model ?model (is ?result (clear-model))))
      ((tell ?form) :-
       (model-name ?model)
       (in-model ?model (is ?result ?form)))
      ((ask-in-model ?question) :-
       (model-name ?model)
       (in-model ?model ?question))))

(model-object-database modelled-alison-object-database
  :built-on 'modelled-person-object-database
  :clauses
    '(((model-name "ALISON-MODEL"))))

(model-object-database modelled-room-object-database
  :built-on 'modelled-person-object-database
  :clauses
    '(((model-name "ROOM-MODEL"))))

(model-object-database searle-object-database
  :built-on 'modelled-person-object-database
  :clauses
    '(((model-name "SEARLE-MODEL"))

      ((do-turing-test) :-
       (do-scripts 1))

      ((do-scripts 10) :-
       (do-script 10))
      ((do-scripts ?count) :-
       (call-without-result (do-script ?count))
       (is ?newcount (+ '?count 1))
       (do-scripts ?newcount))

      ((do-script ?count) :-
       (nl) (nl)
       (write-list (writing script number ?count))
       (get-script ?script)
       (= (?statements ?questions) ?script)
       (self ?self)
       (tell-script-statements ?self ?statements)
       (check-script alison ?statements ?questions)
       (check-script chinese-room ?statements ?questions))

      ((check-script ?subject ?statements ?questions) :-
       (nl) (nl)
       (write-list (telling the script to ?subject))
       (tell-script-statements ?subject ?statements)
       (nl) (nl)
       (write-list (now asking questions))
```

```
 (evaluate-answers ?subject ?questions))

((tell-script-statements ?subject ?statements) :-
 (is ?result (ask-object-to ?subject (clear)))
 (tell-statements ?subject ?statements))

((tell-statement ?subject ?statement) :-
 (self ?subject)
 (write-list (?statement))
 (ask-object-to ?subject (tell ?statement)))
((tell-statement ?subject ?statement) :-
 (ask-object-to ?subject (tell ?statement)))

((tell-statements ?subject ()))
((tell-statements ?subject (?statement . ?statements)) :-
 (cut)
 (tell-statement ?subject ?statement)
 (tell-statements ?subject ?statements))

((ask-question ?subject ?question) :-
 (self ?self)
 (write-list (ask-in-model ?subject ?question))
 (ask-object-to ?subject (ask-in-model (ask-object-to alison
                                         (handle-question ?subject-answer ?question))))
 (write-list (my answer))
 (ask-object-to ?self (ask-in-model (ask-object-to alison
                                      (handle-question ?known-answer ?question))))
 (= ?subject-answer ?known-answer)
 (write-list (the answer is correct)))
((ask-question ?subject ?question) :-
 (write-list (the answer is not correct)))

((evaluate-answers ?subject ?questions) :-
 (ask-and-rate-questions ?subject ?questions ?subject-total ?subject-correct)
 (is ?rating (float (/ '?subject-correct '?subject-total)))
 (is ?form (object-form (find-object '?subject)))
 (write-list (rating for ?subject of form ?form is ?rating))
 (add-rating intentional ?subject ?form ?rating))

((ask-and-rate-questions ?subject () 0 0))
((ask-and-rate-questions ?subject (?question . ?questions) ^newtotal ^newcorrect) :-
 (ask-question ?subject ?question)
 (ask-and-rate-questions ?subject ?questions ?total ?correct)
 (is ?newcorrect (+ 1 '?correct))
 (is ?newtotal (+ 1 '?total)))
((ask-and-rate-questions ?subject (?question . ?questions) ^newtotal ^correct) :-
 (ask-and-rate-questions ?subject ?questions ?total ?correct)
 (is ?newtotal (+ 1 '?total)))

((result no ?stance-to ?action ?situation ^situation) :-
 (self ?self)
 (write-list (?self cant take any stance to ?stance-to for))
 (nl) (spaces 2) (write ?action)
 )))
```

similarity.factor

```
;;; Similarity code derived from various principles and factors derived from the procedures of
;;; numerical taxonomy.  We are using this to provide some kind of useful comparison between the
;;; different physical forms involved.
;;;
;;; We use the Rogers-Tanimoto coefficient to compare the single state features.  Note this does not
;;; have negative matches.
```

```
(defun rogers-tanimoto-coefficient (form1 form2)
  (let ((characters (array-dimension form1 0))
        (matched 0)
        (unmatched 0))
    (loop for character from 0 below characters
          do (if (= (aref form1 character) (aref form2 character))
                 (incf matched)
                 (incf unmatched)))
    (/ (- matched unmatched) (+ matched unmatched))))
```

;;; The correlation coefficient is one of the best metrics, in that it probably comes the closest to the
;;; classification that people produce. For that reason, we will focus on it as a metric to be used in
;;; our measurement of similarity. This is the product moment correlation coefficient.

```
(defun sqr (x)
  (* x x))

(defun correlation-coefficient (form1 form2)
  (when (equalp form1 form2)
    (return-from correlation-coefficient 1.0))
  (let* ((characters (array-dimension form1 0))
         (mean1 (/ (loop for character from 0 below characters
                         sum (aref form1 character))
                   characters))
         (mean2 (/ (loop for character from 0 below characters
                         sum (aref form2 character))
                   characters))
         (numerator (loop for character from 0 below characters
                          sum (* (- (aref form1 character) mean1)
                                 (- (aref form2 character) mean2))))
         (denominator (sqrt (* (loop for character from 0 below characters
                                     sum (sqr (- (aref form1 character) mean1)))
                               (loop for character from 0 below characters
                                     sum (sqr (- (aref form2 character) mean2)))))))
    (if (zerop denominator)
        (signum numerator)
        (/ numerator denominator))))
```

;;; The similarity metric here is normalised to a value of between zero (dissimilar) and one (identical,
;;; as far as this metric is concerned). We should keep to a linear rating to accommodate the results
;;; of Eddy *et al.*

```
(defun get-similarity (form1 form2)
  (ensure-forms)
  (let* ((binary-form1 (get-similarity-table form1 'similarity-binary-values))
         (binary-form2 (get-similarity-table form2 'similarity-binary-values))
         (real-form1 (get-similarity-table form1 'similarity-real-values))
         (real-form2 (get-similarity-table form2 'similarity-real-values))
         (binary-measure (rogers-tanimoto-coefficient binary-form1 binary-form2))
         (real-measure (correlation-coefficient real-form1 real-form2)))
    (sqr (/ (+ binary-measure (/ (+ 1 real-measure) 2)) 2))))
```

familiarity.factor

;;; The familiarity component. This associates values with forms, so that we associate different
;;; classes of stance with different behaviours. The idea is to promote the kinds of transfer by
;;; similarity of form due to familiarity.
;;;
;;; The inputs are a form, a stance, and a rating between −1 and +1. We build up a rating which, for
;;; a given form, allows us to get out a rating for each stance. Use *add-rating* to record an object's

;;; rating. Then we can use this to derive a familiarity rating to affect our ascription of mentality to
;;; the room. This then leaves us clear to implement the complex factor, in which Searle reduces the
;;; ascription of mentality when aware of the constitutional structure of the Chinese Room.

```
(defpredicate add-rating
  ((add-rating ?stance ?individual ?form ?rating) :-
   (familiarity ?values)
   (retract ((familiarity ?values)))
   (asserta ((familiarity ((?stance ?individual ?form ?rating) . ?values)))))
  ((add-rating ?stance ?individual ?form ?rating) :-
   (asserta ((familiarity ((?stance ?individual ?form ?rating)))))))

(defpredicate get-familiarity
  ((get-familiarity ?stance ?individual ?form 0.0) :-
   (familiarity ()))
  ((get-familiarity ?stance ?individual ?form ^rating) :-
   (familiarity ?values)
   (get-familiarity-internal ?stance ?individual ?form ?values 1.0 0.0 0.0 ?rating))
  ((get-familiarity ?stance ?individual ?form 0.0)))

(defconstant familiarity-decay 0.8)
(defconstant individual-factor 0.7)

(defun familiarity-function (this-individual this-form individual form rating)
  (+ (if (eq this-individual individual)
         (* individual-factor rating)
         0.0)
     (* (- 1 individual-factor) (get-similarity this-form form) rating)))

(defun familiarity-rating (total divisor)
  (* (/ total divisor)
     (expt 2 (* -5 (/ (- (1+ (/ familiarity-decay (- 1 familiarity-decay)))
                         divisor)
                      (/ familiarity-decay (- 1 familiarity-decay)))))))

(defpredicate get-familiarity-internal
  ((get-familiarity-internal ?stance ?individual ?form
     () ?multiplier ?total ?divisor ^result) :-
   (is ?result (familiarity-rating '?total '?divisor)))
  ((get-familiarity-internal ?stance ?individual ?form
     ((?stance ?other-individual ?other-form ?rating) . ?rest) ?multiplier ?total
     ?divisor ^result) :-
   (is ?new-multiplier (* '?multiplier familiarity-decay))
   (is ?new-divisor (+ '?divisor '?multiplier))
   (is ?new-total (+ '?total (* (familiarity-function
                                  '?individual '?form
                                  '?other-individual '?other-form '?rating)
                                '?multiplier)))
   (get-familiarity-internal ?stance ?individual ?form
     ?rest ?new-multiplier ?new-total ?new-divisor ^result))
  ((get-familiarity-internal ?stance ?individual ?form
     (?first . ?rest) ?multiplier ?total ?divisor ?result) :-
   (get-familiarity-internal ?stance ?individual ?form
     ?rest ?multiplier ?total ?divisor ^result)))

(defun get-familiarity (from-name stance to-name to-form)
  (let ((bindings (ask-object-for-bindings from-name
                    `(get-familiarity ,stance ,to-name ,to-form ?rating))))
    (instantiate '?rating (first bindings))))
```

animation.factor

;;; We use only a very simple implementation of the animation factor. For this, we use the form
;;; information that has also been attached to the form definitions used for the similarity
;;; measurement. The difference is that we use the animation information rather than the similarity
;;; information.

```
(defun get-animation (form)
  (if (get-form-character-value form 'animated?)
      1.0
      0.0))
```

structure.factor

;;; We're getting to the home straight now. There is just the one more factor that we're going to
;;; implement completely. This is the one involving structure. We have a couple of effects to capture
;;; here. One is the apparent level system, which means that Searle will try to identify with Searle-in-
;;; the-room rather than the room, because Searle-in-the-room in more similar to Searle than the
;;; room is as a whole. The other effect is that the more that is known about the structure of a
;;; system, the less that will be ascribed to the system as a whole.
;;;
;;; The second effect is really quite easy. We can use the inside relation to generate all the (known)
;;; internals of a system and turn that into a rating about the ascription. This will generally be quite
;;; dramatic: that is, only one or two elements will have a significant effect, where more elements will
;;; tend to tail off the values.
;;;
;;; The first effect appears to be a transferred ascription. When it is difficult to ascribe mentality to
;;; the whole, we can ascribe it to constituent elements, when they are compatible. A change in form
;;; will affect this ascription. In fact, it will also affect the other effect, although perhaps for a
;;; different reason.

```
(defpredicate find-all-contained-structures
  ((find-all-contained-structures ?believer ?object ^complete-result) :-
   (get-situation ?situation)
   (find-all-contained-structures-in-situation ?believer ?situation ?object ?result)
   (difference ?result (?object) ?complete-result)))

(defpredicate find-all-contained-structures-in-situation
  ((find-all-contained-structures-in-situation ?believer () ?object ()))
  ((find-all-contained-structures-in-situation ?believer ?situation ?object ^result) :-
   (search-situation-structures ?believer ?situation (?object) ?result)))

(defpredicate search-situation-structures
  ((search-situation-structures ?believer ?situation () ()))
  ((search-situation-structures ?believer ?situation
     (?object . ?rest) (^object . ^result)) :-
   (find-contained-structures-in-situation ?believer ?situation ?object ?values)
   (append ?values ?rest ?new-rest)
   (search-situation-structures ?believer ?situation ?new-rest ?result)))

(defpredicate find-contained-structures-in-situation
  ((find-contained-structures-in-situation ?believer () ?object ()))
  ((find-contained-structures-in-situation ?believer
     ((believes ?believer (inside ?something ?object)) . ?situation-rest)
     ?object (^something . ^rest)) :-
   (find-contained-structures-in-situation ?believer ?situation-rest ?object ?rest))
  ((find-contained-structures-in-situation ?believer
     ((believes ?someone (inside ?something ?other-object)) . ?situation-rest)
     ?object ^rest) :-
   (find-contained-structures-in-situation ?believer ?situation-rest ?object ?rest)))
```

```
;;; The length of this list will affect the ascription, but so will a rating of each individual element.  We
;;; can use the similarity metric again for this.  The goal should be for this factor to be relatively high
;;; for a well known structure, or one with just a few elements in which happen to be rather like the
;;; ascriber.

(defun calculate-structure-rating (form forms)
  (when (null forms)
    (return-from calculate-structure-rating 0.0))
  (* (/ 1 (length forms))
     (loop for other-form in forms
           sum (get-similarity form other-form) into total
           finally do (return (/ total (length forms))))))

;;; There are two factors here: first, a calculation on the number of forms which go to make up this
;;; structure, and second, the mean similarity of those forms.

(defun get-structure (from-name to-name)
  (let* ((bindings (ask-object-for-bindings from-name
                     `(find-all-contained-structures ,from-name ,to-name ?structure)))
         (structure (instantiate '?structure (first bindings))))
    (calculate-structure-rating
      (object-form (find-object from-name))
      (loop for object in structure
            collect (object-form (find-object object))))))
```

predicate.factor

```
;;; A simple implementation of the predicate effects.  This simply classifies predicates as either
;;; physical or intentional.

(defun get-intentional-predicate (predicate)
  (if (member predicate '(perceived believes))
      1.0
      0.0))

(defun get-physical-predicate (predicate)
  (if (member predicate '(put-in take-out place))
      1.0
      0.0))
```

Appendix D
Model program listings

run.def

```
;;; -*- Mode: Lisp; Package: COMMON-LISP-USER -*-
;;;
;;; Seeing things as people: anthropomorphism in common-sense psychology
;;; Thesis models: load this file to run all the models in turn.
;;;
;;; First define the model package, and a few auxiliary functions that allow us to load and include
;;; the other files that go to make up the models.

(defpackage "MODEL"
  (:use "COMMON-LISP"))

(in-package "MODEL")

(assert *load-truename* ()
  "These models can only be run by loading this file.")

(defparameter *program-directory* (directory-namestring *load-truename*))
(defvar *loaded-files* ())
(defvar *load-depth* 0)

(defun load-file (pathname)
  (format *standard-output* "~&; ~v@TLoading file ~S..."
    *load-depth* (write-to-string pathname :escape nil))
  (setf pathname (merge-pathnames pathname *program-directory*))
  (let ((*load-depth* (1+ *load-depth*)))
    (load pathname :verbose nil))
  (push pathname *loaded-files*)
  pathname)

(defun include-file (pathname)
  (format *standard-output* "~&; ~v@TIncluding file ~S..."
    *load-depth* (write-to-string pathname :escape nil))
  (setf pathname (merge-pathnames pathname *program-directory*))
  (with-open-file (input (merge-pathnames pathname *program-directory*) :direction :input)
    (loop for form = (read input () '#1=#:end)
          until (eq form '#1#)
          collect form)))
```

;;; Load a general purpose Prolog interpreter, written in Lisp. This interpreter has a Lisp-like syntax
;;; but apart from that, it behaves as a standard Prolog. Its implementation will not be described in
;;; any more detail here.

```
(load-file "prolog.lisp")
```

;;; Load the definition files which include the main model programs. These will all be listed later in
;;; this appendix.

```
(load-file "general-primitives.def")
(load-file "named-objects.def")
(load-file "traces.def")
(load-file "databases.def")
(load-file "object-databases.def")
(load-file "models.def")
(load-file "objects.def")
(load-file "stances.def")
(load-file "model-primitives.def")
(load-file "define-forms.def")
```

;;; Define the *run-model* function. This function loads and runs a file containing one of the thesis
;;; models, and writes its trace output to another file. These model and trace files are shown in
;;; Appendices A and B.

```
(defun run-model (file-name)
  (let ((input (merge-pathnames (make-pathname :type "model") file-name))
        (output (merge-pathnames (merge-pathnames (make-pathname :type "trace") file-name)
                  *program-directory*)))
    (with-open-file (output output :direction :output :if-exists :supersede)
      (let ((*standard-output* (make-broadcast-stream output *terminal-io*)))
        (declare (special *standard-output*))
        (load-file input)))))
```

;;; Run the model files for the standard version of Baron-Cohen *et al*'s false belief test, and its
;;; variations, described in chapters 9 and 10.

```
(run-model "mfb")
(run-model "mfbsimulationfail")
(run-model "mfbsimulation")
(run-model "mfbcopy")
(run-model "mfbsituationfail")
(run-model "mfbsituation")
```

;;; Run the model files for Searle's Chinese Room, and its variations, described in chapter 12.

```
(run-model "mcr")
(run-model "mcrsystems")
(run-model "mcrrobot")
(run-model "mcrneural")
```

general-primitives.def

;;; Some general purpose primitives and predicates added to the Prolog interpreter, for list
;;; processing, generating, random numbers, calling Lisp functions, and so on.

```
(defpredicate member
  ((member ?x (?x . ?z)))
  ((member ?x (?y . ?z)) :-
   (member ?x ?z)))

(defpredicate delete
  ((delete ?x (?x . ?rest) ^rest))
  ((delete ?x (?y . ?rest) (^y . ^rest1)) :-
   (delete ?x ?rest ?rest1)))

(defpredicate difference
  ((difference () ?x ()))
  ((difference ?x () ^x))
  ((difference (?x . ?rest) ?list (^x . ^rest1)) :-
   (not (member ?x ?list))
   (difference ?rest ?list ?rest1))
  ((difference (?x . ?rest) ?list ^rest1) :-
   (difference ?rest ?list ?rest1)))
```

```
(defpredicate append
  ((append () ?x ^x))
  ((append (?x . ?rest1) ?rest2 (^x . ^rest3)) :-
   (append ?rest1 ?rest2 ?rest3)))

(defprimitive those (environment pattern list result)
  (setf list (instantiate list environment))
  (loop for element in list
        for match = (unify pattern element environment)
        if (not (failedp match))
          collect (instantiate pattern match) into patterns
        end
        finally
        do (return (let ((result (unify result patterns environment)))
                     (if (failedp result)
                         nil
                         (list result))))))

(defprimitive requote (environment pattern list result-pattern result)
  (setf list (instantiate list environment))
  (loop for element in list
        for match = (unify pattern element environment)
        if (not (failedp match))
          collect (instantiate result-pattern match) into patterns
        end
        finally
        do (return (let ((result (unify result patterns environment)))
                     (if (failedp result)
                         nil
                         (list result))))))

(defprimitive call-without-result (environment query)
  (let ((results (prove query :environment environment)))
    (if (null results)
        ()
        (list environment))))

(defprimitive when (environment function &rest values)
  (if (apply (symbol-function function) (instantiate values environment))
      (list environment)
      ()))

(defprimitive call (environment function &rest values)
  (apply (symbol-function function) (instantiate values environment))
  (list environment))

(defprimitive length (environment variable list)
  (unified-environments variable (length (instantiate list environment))
    environment))

(defprimitive random (environment variable lower upper)
  (let ((lower (instantiate lower environment))
        (upper (instantiate upper environment)))
    (unified-environments variable (+ lower (random (1+ (- upper lower))))
      environment)))
```

named-objects.def

```
;;; An abstract class for named objects.  More or less all objects in the modelling system are named,
;;; so this class is used by most other classes in the model framework.

(defclass named-object ()
  ((name
     :initarg :name
     :reader object-name)))
```

traces.def

```
;;; Some extra tracing primitives.  You can use these primitives to choose the level of tracing you
;;; want when you're watching a model being run.

(defvar *trace-model-flags* ())

(defprimitive start-trace (environment keyword)
  (setf keyword (instantiate keyword environment))
  (pushnew keyword *trace-model-flags*)
  (list environment))

(defprimitive stop-trace-all (environment)
  (setf *trace-model-flags* ())
  (list environment))

(defprimitive stop-trace (environment keyword)
  (setf keyword (instantiate keyword environment))
  (setf *trace-model-flags* (delete keyword *trace-model-flags*))
  (list environment))

(defprimitive trace (environment keyword)
  (setf keyword (instantiate keyword environment))
  (if (member keyword *trace-model-flags*)
      (list environment)
      ()))
```

databases.def

```
;;; A general database framework for the Prolog interpreter.  By default, the interpreter does not use
;;; it own database, but uses a 'hook' function to get all the clauses for a given predicate.  Here we
;;; define a database class, which can be mixed into other class to give any instances of those classes
;;; their own databases.

(defvar *database*)

(defclass database ()
  ((clauses
     :initform ()
     :accessor database-clauses)))

(defmethod clear ((database database))
  (with-slots (clauses) database
    (setf clauses nil)))

(defun database-clause-hook (predicate)
  (remove-if-not #'(lambda (clause)
                     (eq (first (first clause)) predicate))
                 (database-clauses *database*)))

;;; Implement database primitives retract, asserta, and assertz.  These can be used to add and remove
;;; clauses from the current database.  The model system uses this principle to give objects models of
;;; the world, and of each other.

(defprimitive retract (environment clause &aux (database *database*))
  (setf clause (instantiate clause environment))
  (with-slots (clauses) database
    (setf clauses
          (delete-if #'(lambda (database-clause)
                         (not (failedp (unify clause database-clause ()))))
                     clauses)))
  (list environment))
```

```
(defprimitive asserta (environment clause &aux (database *database*))
  (setf clause (instantiate clause environment))
  (with-slots (clauses) database
    (push clause clauses))
  (list environment))

(defprimitive assertz (environment clause &aux (database *database*))
  (setf clause (instantiate clause environment))
  (with-slots (clauses) database
    (setf clauses (nconc clauses (list clause))))
  (list environment))
```

object-databases.def

```
;;; Databases to be built on other databases. This is used to avoid having to copy all the common
;;; clauses everywhere. Instead, we can just build on an earlier database and add the rules that have
;;; changed. Do this by adding a database object class.

(defclass object-database (named-object database)
  ((clauses
     :initarg :clauses)
   (built-on
     :initarg :built-on
     :reader object-database-built-on)))

(defmacro model-object-database (name &rest options)
  (if options
      `(make-object-database ',name ,@options)
      `(find-object-database ',name)))

(defun make-object-database (name &key built-on clauses)
  (setf (get name 'object-database) (make-instance 'object-database
                                      :name name
                                      :clauses clauses
                                      :built-on built-on))
  name)

(defun find-object-database (name)
  (let ((object-database (get name 'object-database '#1=#:error)))
    (if (eq object-database '#1#)
        (error "Can't find object database named ~A" name)
        object-database)))

(defmethod database-clauses ((object-database object-database))
  (append (slot-value object-database 'clauses)
          (let ((built-on (object-database-built-on object-database)))
            (when built-on
              (database-clauses (find-object-database built-on))))))
```

models.def

```
;;; Models are objects with their own database. Also, it is possible for more than one model to be
;;; open and running at the same time. Add a model class, and functions to switch between models.

(defvar *models* ())
(defvar *model*)

(defclass model (database named-object)
  ((objects
     :initform nil
     :accessor model-objects)))
```

```
(defmethod clear ((model model))
  (call-next-method)
  (setf (model-objects model) ()))

(defun find-object (name &optional (model *model*))
  (getf (model-objects model) name))

(defun clear-model (&optional (model *model*))
  (clear model))

(defun make-model (name)
  (make-instance 'model :name name))

(defun make-anonymous-model ()
  (make-instance 'model))

(defun find-model (name)
  (find name *models* :test #'string= :key #'object-name))

(defun in-model (name &key clearp)
  (setf *model*
        (or (find-model name)
            (let ((model (make-model name)))
              (push model *models*)
              model)))
  (when clearp (clear-model *model*))
  name)
```

;;; Here are the additional clauses needed by a model. Note that these live outside the main data
;;; base, but are permanently tagged on by the *database-clauses* method. The clauses themselves are
;;; taken from figure 9.1.

```
(defmethod database-clauses ((model model))
  (append (call-next-method)
          '#.(include-file "09.1.physmodel.freehand.text")))

(defun proved (expression)
  (let ((environments (prove expression)))
    (when environments
      (instantiate expression (first environments)))))

(defun ask-model-function (expression)
  (let ((*clause-hook* #'database-clause-hook)
        (*database* *model*))
    (proved expression)))

(defmacro tell-model (expression)
  `(ask-model-function '(tell ,expression)))

(defmacro ask-model (expression)
  `(ask-model-function ',expression))
```

objects.def

;;; Within a model, any named object should be associated with a model description, which will be
;;; instance of the class object. We use another macro to do this, and we make this definition in the
;;; model, where it can be accessed.

```
(defclass object (database named-object)
  ((form
     :initarg :form
     :accessor object-form)
   (stances
```

```
    :initform ()
    :initarg :stances
    :accessor object-stances)
  (taken-stance-to
    :accessor object-taken-stance-to)
  (taken-stance-for
    :accessor object-taken-stance-for)
  (selected-stance
    :initform nil
    :accessor object-selected-stance)
  (database
    :initform ()
    :initarg :database
    :accessor object-database)))

(defmethod database-clauses ((object object))
  (append (call-next-method)
          (when (object-selected-stance object)
            (database-clauses (find-stance (object-selected-stance object))))
          (database-clauses (find-object-database (object-database object)))))

(defun model-object-function (name class &rest options)
  (let ((object (apply #'make-instance class :name name options)))
    (setf (getf (model-objects *model*) name) object)
    name))

(defmacro model-object (name &body options)
  (let ((class (getf options :class 'object))
        (database (getf options :database 'basic-object-database))
        (stances (getf options :stances)))
    (remf options :class)
    (remf options :database)
    (remf options :stances)
    `(model-object-function ',name ',class
      :name ',name
      :database ',database
      :stances (list ,@(mapcar #'(lambda (stance)
                                   (list 'quote stance))
                               stances))
      ,@(loop for rest-options on options by #'cddr
              for keyword = (first rest-options)
              for value = (second rest-options)
              append (list keyword (list 'quote value))))))
```

;;; We can find an object easily enough through this little database. The point is that now we can
;;; handle the *notify-object-event* through finding the object and asking it if it is interested in the
;;; event. The key function here is *notify-object-event*. This can be set up to find the named object
;;; and, when it has been found, to ask it to handle the event. This is where the characteristics of the
;;; individual objects then start to come into their own.

```
(defun notify-object-event (name event)
  (ask-object* name `(handle-event ,event) () 'event))
```

;;; There are two kinds of thing we can ask an object to do, namely an action or a question. These
;;; are handled the same way internally, but we use different functions here so that we can trace them
;;; separately and make the code clearer.

```
(defvar *self*)
```

;;; Whenever an object is processing a model operation, we bind a variable to it so that we can
;;; get this back from the self predicate. This is a way of finding out who oneself is.

```
(defun ask-object* (name expression bindingsp class &key environment (form expression))
  (let* ((*clause-hook* #'database-clause-hook)
         (*database* (find-object name *model*))
         (*self* name))
    (when (member class *trace-model-flags*)
```

```
      (let ((*print-pretty* t)
            (*print-case* :downcase))
        (format t "~&~{~W~^ ~}" `(,class ,name ,form))))
    (if bindingsp
        (prove expression :environment environment)
        (proved expression))))

(defun ask-object-for-bindings (name expression)
  (ask-object* name expression t 'action))

(defun ask-object-to-function (name expression)
  (ask-object* name expression () 'action))

(defmacro ask-object-to (name expression)
  `(ask-object-to-function ',name ',expression))

(defmacro ask-object-if (name expression)
  `(ask-object-if-function ',name ',expression))

(defun ask-object-if-function (name expression)
  (ask-object* name `(handle-question ,(make-symbol "?response") ,expression)
    () 'question :form expression))
```

stances.def

;;; Stances need to be taken with care. An object must adopt a stance to a particular thing, or object,
;;; but the database for the whole object shouldn't need to change. That is, the intentional stance
;;; should need to be a database in its own right. It should add belief rules to the object's database
;;; for its predictions with respect to a particular object. Define a stance as an abstract class. The
;;; different stances will inherit from this.

```
(defclass stance (named-object database)
  ((clauses
     :initarg :clauses)
   (dispositions
     :initarg :dispositions
     :initform nil)
   (disposition-energy
     :initarg :disposition-energy
     :reader stance-disposition-energy)
   (built-on
     :initarg :built-on
     :reader stance-built-on)
   (class
     :initarg :class)))
```

;;; The stance's energy is minus the sum of all the dispositions for the stance. The more favour by
;;; the dispositions, the lower the energy.

```
(defvar self)

(defmethod stance-energy ((from object) to stance)
  (progv (list 'self) (list from)
    (loop for disposition in (stance-dispositions stance)
          sum (funcall-disposition-energy stance disposition
                :from from :to to
                :predicate (object-taken-stance-for from)))))

(defmacro model-stance (name &rest options)
  (if options
      `(make-stance ',name ,@options)
      `(find-stance ',name)))
```

;;; One stance may be built on another. When this happens, clauses should be appended and other
;;; slots combined appropriately.

```
(defmethod stance-dispositions ((stance stance))
  (union (slot-value stance 'dispositions)
         (let ((built-on (stance-built-on stance)))
           (when built-on
             (stance-dispositions (find-stance built-on))))))

(defmethod database-clauses ((stance stance))
  (append (slot-value stance 'clauses)
          (let ((built-on (stance-built-on stance)))
            (when built-on
              (database-clauses (find-stance built-on))))))
```

```
(defmethod stance-class ((stance stance))
  (if (slot-boundp stance 'class)
      (slot-value stance 'class)
      (stance-class (stance-built-on (find-stance stance)))))
```

;;; The generic function used for the disposition energy is a bit tricky to combine, because of
;;; *compute-applicable-methods.*

```
(defmethod funcall-disposition-energy ((stance stance) keyword &rest options)
  (let* ((function (stance-disposition-energy stance))
         (applicablep (and function (compute-applicable-methods function (list keyword)))))
    (if applicablep
        (apply (stance-disposition-energy stance) keyword options)
        (apply #'funcall-disposition-energy
          (find-stance (stance-built-on stance)) keyword options))))

(defun make-stance (name &key built-on clauses dispositions disposition-energy class)
  (setf (get name 'stance) (make-instance 'stance
                             :name name
                             :clauses clauses
                             :class class
                             :dispositions dispositions
                             :disposition-energy disposition-energy
                             :built-on built-on))
  name)

(defun find-stance (name)
  (let ((stance (get name 'stance '#1=#:error)))
    (if (eq stance '#1#)
        (error "Can't find stance named ~A" name)
        stance)))
```

;;; *select-stance* chooses the possible stance with the least energy. This represents the most favoured
;;; stance at this moment in time, to this particular object.

```
(defparameter stance-energy-limit 0.0)

(defmethod select-stance ((object object) to
                          &aux (best-stance nil) (minimum-stance-energy stance-energy-limit))
  (loop for stance in (object-stances object)
        for stance-energy = (stance-energy object to (find-stance stance))
        when (< stance-energy minimum-stance-energy)
        do (setf minimum-stance-energy stance-energy)
           (setf best-stance stance)
        end
        finally do (return best-stance)))
```

model-components.def

```
(defprimitive in-model (environment name query)
  (let ((old-model *model*))
    (unwind-protect (progn
                      (in-model (instantiate name environment))
                      (prove (instantiate query environment) :environment environment))
      (setf *model* old-model))))
```

;;; An object depends on being able to see objects. We do this through the model, and it is just
;;; assumed. This is much more realistic than maintaining any sophisticated structure in the objects
;;; themselves. It is not totally assumed, though, because not everything can see. It depends on
;;; whether or not this primitive is used.

```
(defprimitive sees (environment object variable)
  (let* ((pattern (instantiate `(inside ,object ?x) environment))
         (*clause-hook* #'database-clause-hook)
         (*database* *model*)
         (environments (prove pattern))
         (location (rest (lookup-variable '?x (first environments)))))
    (when location
      (loop with result
            for this-environment in (prove `(inside ?x ,location) :all t)
            for match = (lookup-variable '?x this-environment)
            if match
              do (setq result (unify variable (rest match) environment))
              and if (not (failedp result))
                    collect result
                  end
            end))))

(defprimitive ask-object-to (environment name expression)
  (let* ((name (instantiate name environment))
         (expression (instantiate expression environment)))
    (ask-object* name expression t 'action :environment environment)))

(defprimitive ask-object-if (environment name expression)
  (let* ((name (instantiate name environment))
         (expression (instantiate expression environment)))
    (ask-object* name `(handle-question ,(make-symbol "?response") ,expression) t 'question
      :form expression :environment environment)))

(defprimitive self (environment variable)
  (let ((result (unify variable *self* environment)))
    (if (not (failedp result))
        (list result)
        ())))

(defprimitive in-self (environment self expression)
  (let ((*self* (instantiate self environment)))
    (prove expression :environment environment)))
```

;;; The *in-stance* primitive needs to be handled with care. It should take a stance to something for a
;;; particular prediction and then it should invoke a new proof in that context. On exit the stances
;;; should be restored to their previous context so that processing can continue.

```
(defprimitive in-stance (environment to-pattern for-pattern query-pattern)
  (let* ((to (instantiate to-pattern environment))
         (for (instantiate for-pattern environment))
         (query (instantiate query-pattern environment))
         (object *database*)
         (to-object (find-object to)))
    (assert (and (not (variablep to)) (not (variablep for))))
    (setf (object-taken-stance-to object) to)
    (setf (object-taken-stance-for object) for)
    (let ((stance (select-stance object to-object))
          (old-stance (object-selected-stance object)))
```

```
(unwind-protect (progn
                  (setf (object-selected-stance object) stance)
                  (prove query :environment environment :all nil))
  (setf (object-selected-stance object) old-stance)))))
```

define-forms.def

;;; Similarity code derived from various principles and factors derived from the procedures of
;;; numerical taxonomy. We are using this to provide some kind of useful comparison between the
;;; different physical forms involved. Based on formulae from Sokal and Sneath (1963).
;;;
;;; The *define-form* macro. It includes the macro and all the code that is needed to get everything
;;; ready for the cache used by the similarity metric function.

```
(defparameter *forms-cached-p* ())
(defparameter *forms* ())
(defparameter *form-properties* ())

(defun clear-forms ()
  (setf *forms* ())
  (setf *form-properties* ())
  (setf *forms-cached-p* ()))

(defstruct form
  name
  characters)

(defstruct form-property
  name
  options)

(defmacro get-form (name)
  `(get ,name 'form))
(defmacro get-form-property (name)
  `(get ,name 'form-property))

(defmacro define-form (name characters)
  `(progn
     (setf (get-form ',name) (make-form :name ',name :characters ',characters))
     (pushnew ',name *forms*)
     (setf *forms-cached-p* nil)
     ',name))

(defmacro define-form-property (name &rest options)
  `(progn
     (setf (get-form-property ',name) (make-form-property :name ',name :options ',options))
     (pushnew ',name *form-properties*)
     ',name))

(defun get-form-character-value (name character)
  (let ((character (assoc character (form-characters (get-form name)))))
    (getf (rest character) :value)))

(defun get-form-property-option (name option)
  (let ((value (assoc option (form-property-options (get-form-property name)))))
    (if value
        (second value)
        (error "Can't find form property option ~A" option))))
```

;;; The next stage is to define the code that rebuilds the cache of values used by the similarity system.
;;; This might seem rather tedious, but it also normalises all the values used by the various forms, so

```
;;; that the correlation coefficient works correctly.  See Sokan and Sneath (1963) for details.

(defun get-similarity-characters ()
  (loop for form in *forms*
        with characters = ()
        do (setf characters (union characters
                                   (loop for character in (form-characters (get-form form))
                                         for name = (first character)
                                         when (eq (get-form-property-option name :category)
                                                  'similarity)
                                         collect name)))
        finally do (return characters)))

;;; Now we are able to handle the rest of the system.  When we have finished all the forms, we can
;;; normalise all the data and check that all the types match.  This means we can build two vectors of
;;; binary and real values.  Then we can normalise the real values.  This gives us all we need for a
;;; single function entry point we can use to get the two parts of the similarity metric and combine
;;; them to generate a complete similarity value.

(defmacro get-similarity-table (form type)
  `(get ,form ,type))

(defun build-similarity-tables ()
  (let ((characters (get-similarity-characters)))
    (loop for form in *forms*
          do
            (loop for character in characters
                  for value = (get-form-character-value form character)
                  for type = (get-form-property-option character :type)
                  for result = (ccase type
                                 ((t) (if value 1.0 0.0))
                                 ((real) value))
                  if (eq type 't)
                    collect result into binary-values
                  else
                    collect result into real-values
                  end
                  finally
                  do
                    (setf (get-similarity-table form 'similarity-binary-values)
                          (apply #'vector binary-values))
                    (setf (get-similarity-table form 'similarity-real-values)
                          (apply #'vector real-values))))))

(defun ensure-forms ()
  (or *forms-cached-p*
      (progn
        (build-similarity-tables)
        (setf *forms-cached-p* t)))
  (values))
```